Next Generation Sequencing (NGS) Data Analysis on the HPRC Ada Cluster

Michael Dickens
Texas A&M University
High Performance Research Computing

March 22, 2017

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

Ada Login

- Login to your ada account with X11 forwarding enabled
  - `ssh -X ada.tamu.edu`
  - enable X11 forwarding if using MobaXterm
- Your login password
  - Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts;

Where to Find NGS Tools

- TAMU HPRC Documentation
  - https://hprc.tamu.edu/wiki/index.php/Ada:Bioinformatics
- Type the following UNIX **commands** to see which tools are already installed on Ada
  - `module avail`
  - `module spider toolname` (not case sensitive, but read entire output)
  - `module key assembly` (some modules may be missed because this searches tool descriptions)
- If you find a tool that you want installed on Ada, send an email with the URL link to: **help@hprc.tamu.edu**
  - SeqAnswers http://seqanswers.com/wiki/Software/list
  - omictools.com
- slideshare.net – find shared NGS presentations

Ada Software Toolchains

- Use the same toolchains in your job scripts
  - The **`intel-2015B`** is the recommended toolchain since most bioinformatics tools are installed with it

```
module load Bowtie2/2.2.6-intel-2015B
module load TopHat/2.1.0-intel-2015B
module load Cufflinks/2.2.1-intel-2015B
```

- Avoid loading mixed toolchains:

```
module load Bowtie2/2.2.2-ictce-6.3.5
module load TopHat/2.0.14-goolf-1.7.20
module load Cufflinks/2.2.1-intel-2015B
```

- Avoid loading defaults which may have different toolchains

```
module load Bowtie2 TopHat Cufflinks
```

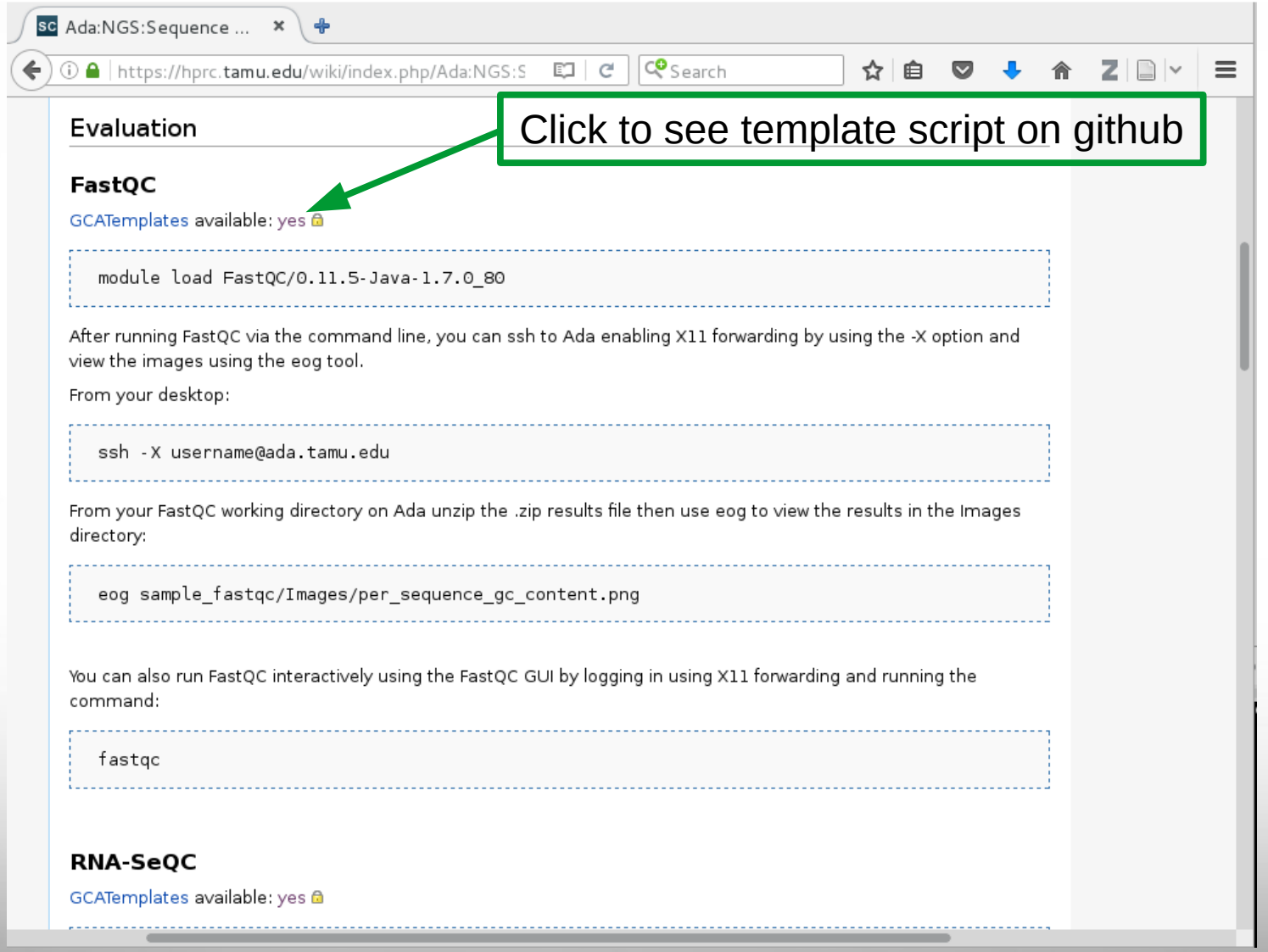
Use $TMPDIR whenever possible

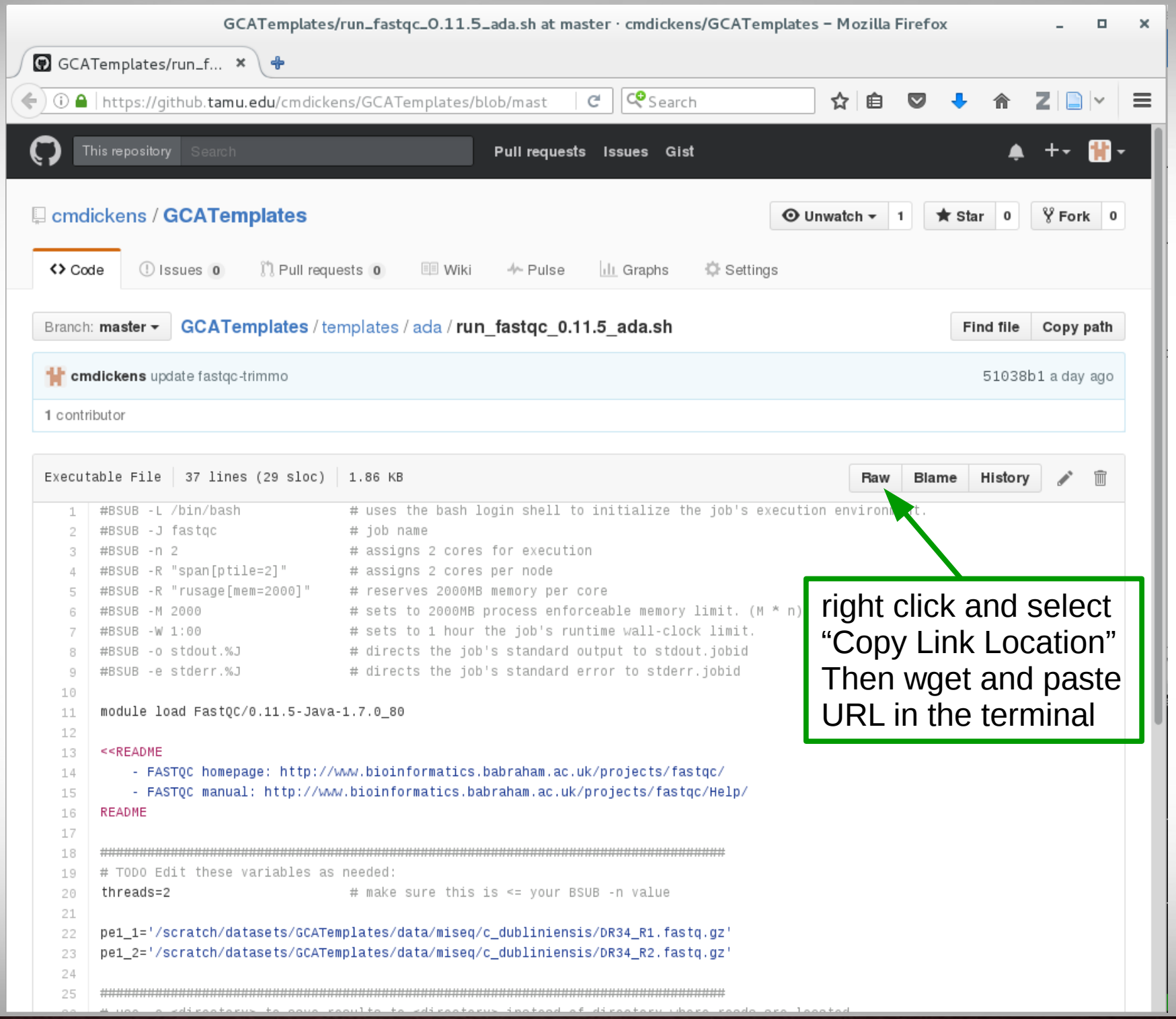
- Use the $TMPDIR if the application you are running can utilize a temporary directory for writing temporary files
- A temp directory ($TMPDIR) is automatically assigned for each job which uses the disk(s) on the compute node (not the $SCRATCH shared file system)
  - Especially useful when a computational tool writes tens of thousands of temporary files which are deleted when the job is finished and are not needed for the final results *
  - This is useful since files on the $TMPDIR will not count against your file quota
  - Don't use $TMPDIR if your software uses temporary files for restarting where it left off if it should stop before completion

```
java -Xmx53g -jar $EBROOTPICARD/FastqToSam.jar TMP_DIR=$TMPDIR \
FASTQ=$pe1_1 FASTQ2=$pe1_2 OUTPUT=$outfile SAMPLE_NAME=$sample_name \
SORT_ORDER=$sort_order MAX_RECORDS_IN_RAM='null'
```

Access GCATemplate Scripts for Ada from the HPRC wiki

https://hprc.tamu.edu/wiki/index.php/Ada:NGS:Sequence_QC#FastQC




GCATemplates/run_fastqc_0.11.5_ada.sh at master · cmdickens/GCATemplates – Mozilla Firefox
https://github.tamu.edu/cmdickens/GCATemplates/blob/mast
Search
This repository
Search
Pull requests
Issues
Gist
cmdickens / GCATemplates
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Settings
Branch: master
GCATemplates / templates / ada / run_fastqc_0.11.5_ada.sh
Find file
Copy path
cmdickens update fastqc-trimmo
51038b1 a day ago
1 contributor
Executable File
37 lines (29 sloc)
1.86 KB
Raw
Blame
History
#BSUB -L /bin/bash # uses the bash login shell to initialize the job's execution environment.
#BSUB -J fastqc # job name
#BSUB -n 2 # assigns 2 cores for execution
#BSUB -R "span[ptile=2]" # assigns 2 cores per node
#BSUB -R "rusage[mem=2000]" # reserves 2000MB memory per core
#BSUB -M 2000 # sets to 2000MB process enforceable memory limit. (M * n)
#BSUB -W 1:00 # sets to 1 hour the job's runtime wall-clock limit.
#BSUB -o stdout.%J # directs the job's standard output to stdout.jobid
#BSUB -e stderr.%J # directs the job's standard error to stderr.jobid
module load FastQC/0.11.5-Java-1.7.0_80
<<README
- FASTQC homepage: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- FASTQC manual: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/
README
# TODO Edit these variables as needed:
threads=2 # make sure this is <= your BSUB -n value
pe1_1='/scratch/datasets/GCATemplates/data/miseq/c_dubliniensis/DR34_R1.fastq.gz'
pe1_2='/scratch/datasets/GCATemplates/data/miseq/c_dubliniensis/DR34_R2.fastq.gz'
right click and select
“Copy Link Location”
Then wget and paste
URL in the terminal

Finding NGS job template scripts using GCATemplates on Ada

```
mkdir $SCRATCH/ngs_class
```

```
cd $SCRATCH/ngs_class
```

```
module load GCATemplates
```

```
gcatemplates
```

Final step will save a template job script file to your current working directory

For practice, we will copy 1 template file

Select #8 then find the template that contains fastqc

After you save the template file:

```
module purge
```

Genomic Computational Analysis Templates

Sample GCATemplate Job Script (Ada)

```
#BSUB -L /bin/bash
#BSUB -J blastx
#BSUB -n 2
#BSUB -R "span[ptile=2]"
#BSUB -R "rusage[mem=2500]"
#BSUB -M 2500
#BSUB -W 2:00
#BSUB -o stdout.%J
#BSUB -e stderr.%J

module load BLAST+/2.2.31-intel-2015B-Python-3.4.3

<<README
    BLAST manual: http://www.ncbi.nlm.nih.gov/books/NBK279690/
README

#blastx: search protein databases using a translated nucleotide query

blastx -query mrna_seqs_nt.fasta -db /scratch/datasets/blast/nr \
-outfmt 10 -out mrna_seqs_nt_blastout.csv
```

Sample GCATemplate Job Script (Ada)

```
#BSUB -L /bin/bash
#BSUB -J blastx
#BSUB -n 2
#BSUB -R "span[ptile=2]"
#BSUB -R "rusage[mem=2500]"
#BSUB -M 2500
#BSUB -W 2:00
#BSUB -o stdout.%J
#BSUB -e stderr.%J

module load BLAST+/2.2.31-intel-2015B-Python-3.4.3

<<README
    BLAST manual: http://www.ncbi.nlm.nih.gov/books/NBK279690/
README

#blastx: search protein databases using a translated nucleotide query

blastx -query mrna_seqs_nt.fasta -db /scratch/datasets/blast/nr \
-outfmt 10 -out mrna_seqs_nt_blastout.csv
```

These parameters are read by the job scheduler

Load the required module(s) first

This is a section of comments

This is a single line comment and not run as part of the script

This is the command to run the application

This means the command is continued on the next line;
The space before the \ is required
Do not put a space after the \

Next Generation Sequencing (NGS)

Illumina Sequencing Technology

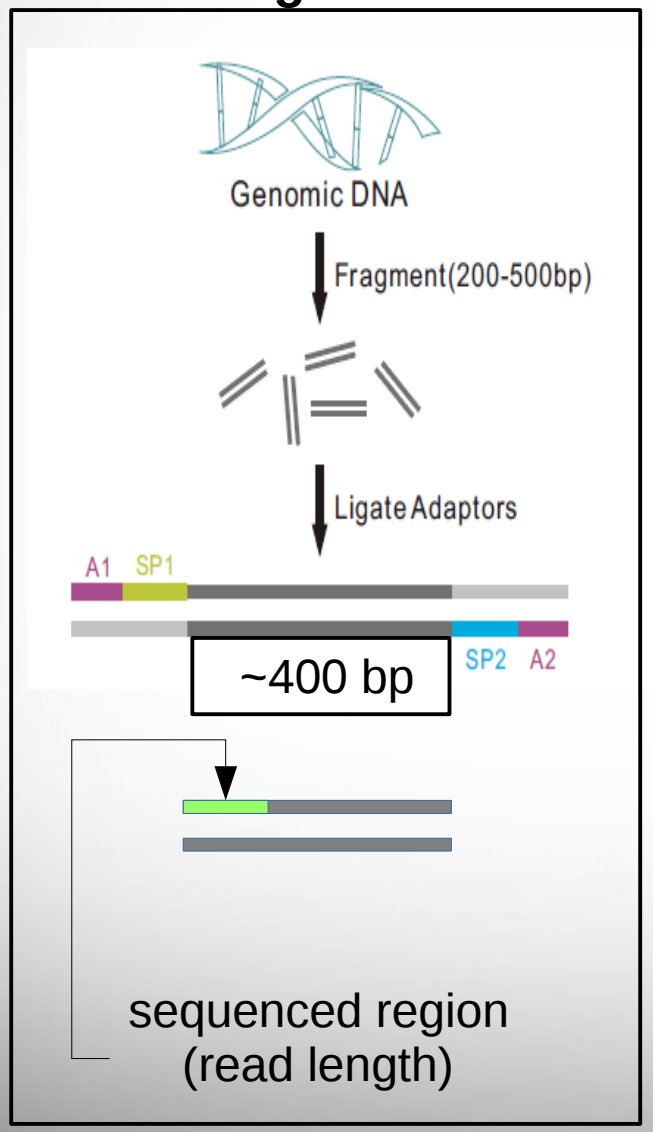
	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No

http://www.illumina.com/systems/sequencing.html (Feb 2017)

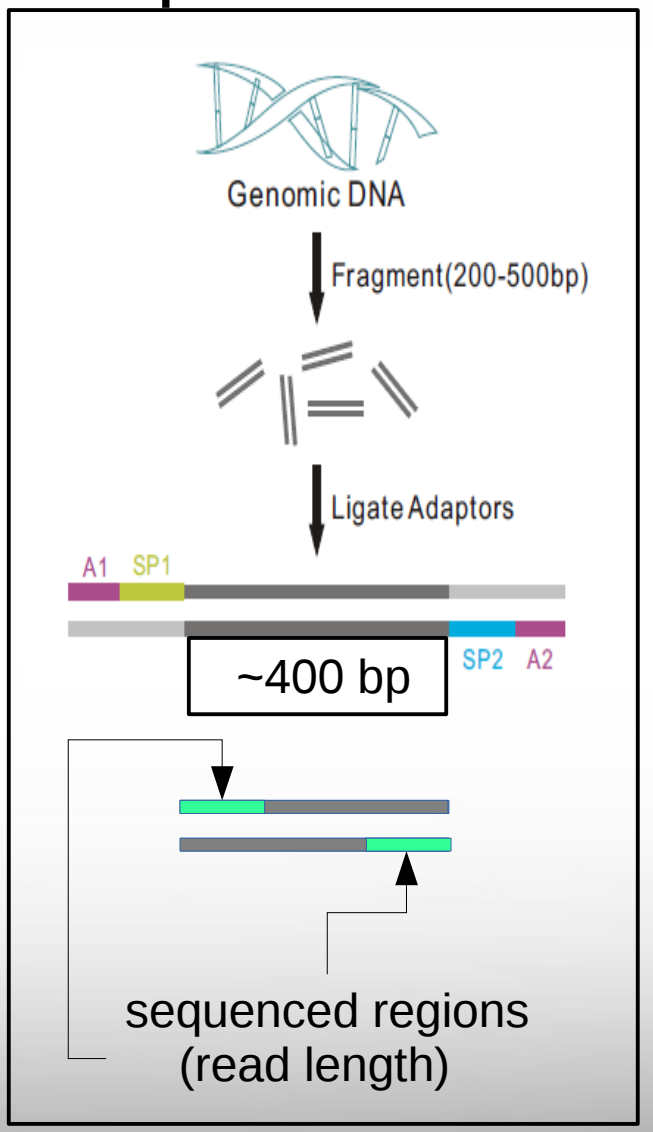
Illumina Sequencing Libraries

illumina.com

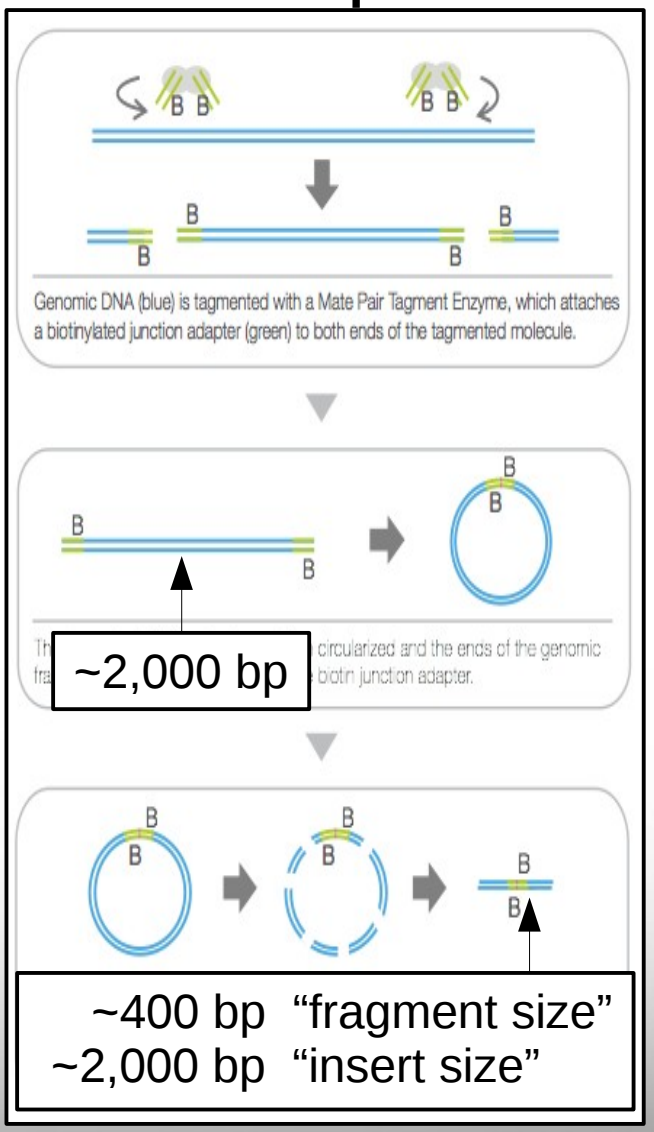
single end



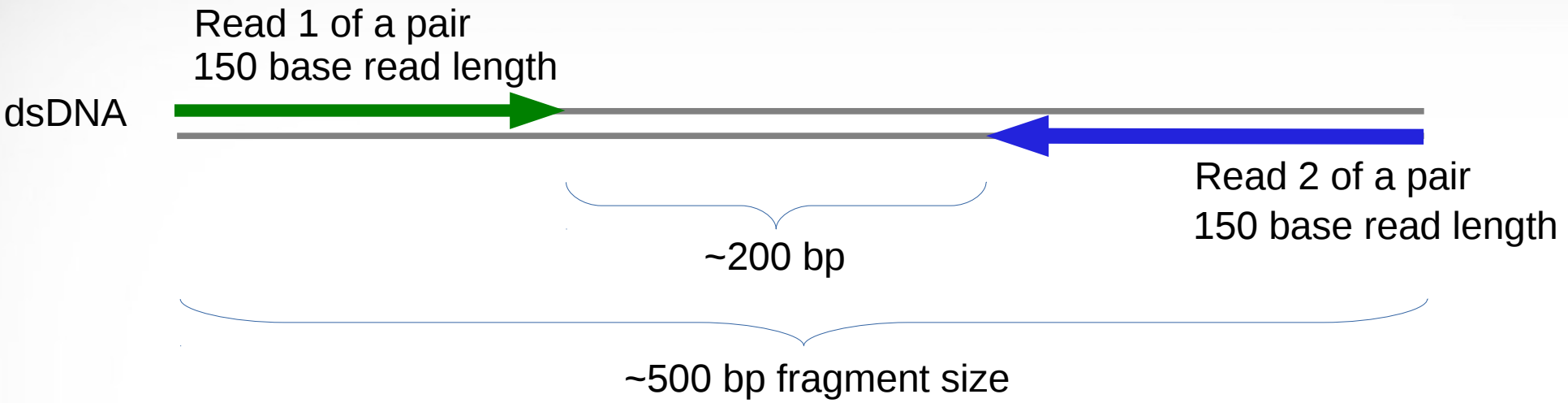
paired ends



mate pairs



Paired End Reads



FASTQ format

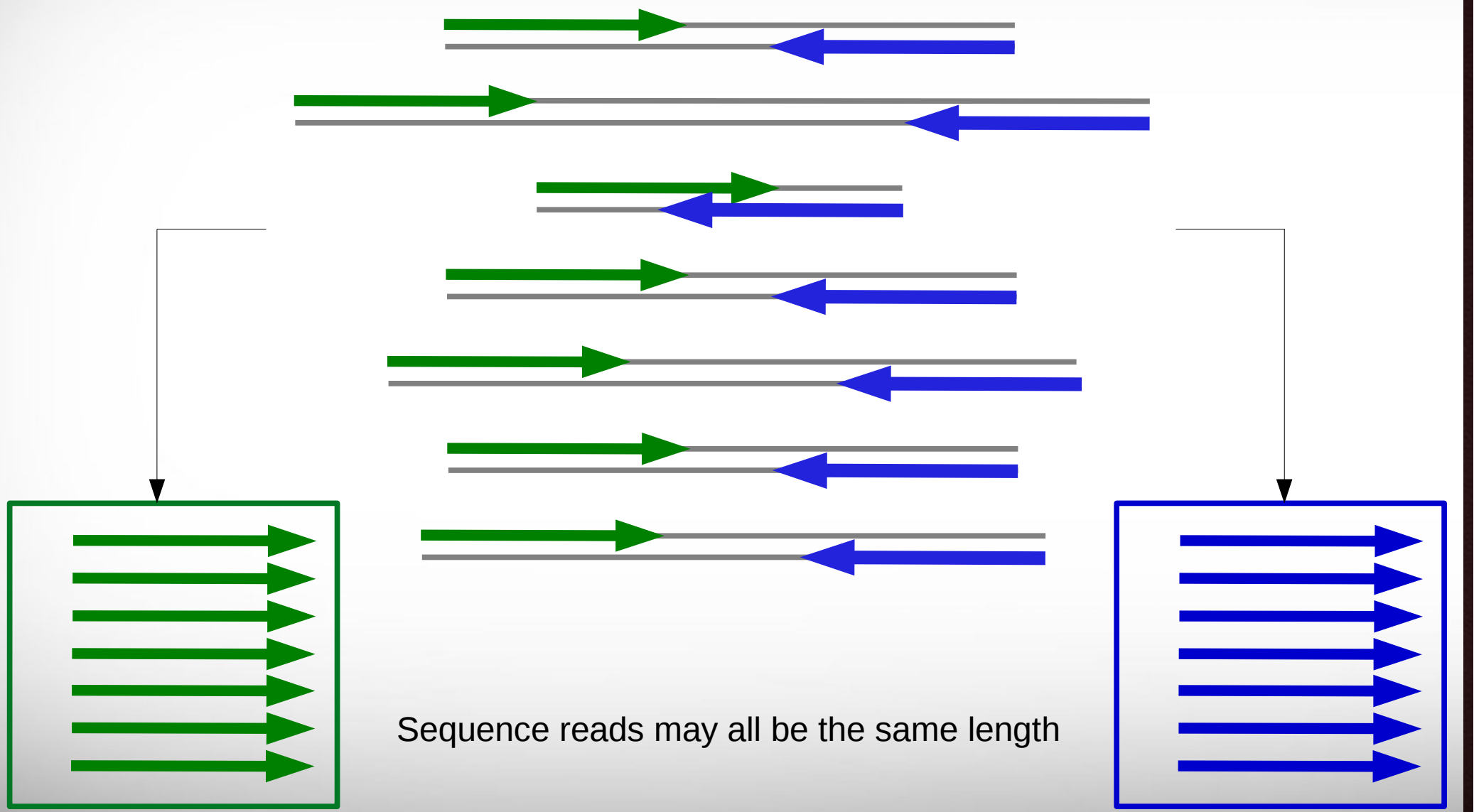
Read 1 pair fastq file

```
@M00861:1:000000000-A36BE:1:1101:14650:1529 1:N:0:8
TTCTTAAAAATACCATAAAAGGCTTAAACTTGCCATTTACGACGGATTAATTCCAACTCTTTTCGGCTATCTTCATCTTTTAAGGTTAAATGACTCATAA
+
FFFHBFFHHIIIIIIHFHHCGEFGHHIHHHIHD/?DGGHHH@DEB,5EGHGHHIIHIF?FGGHHCCBFDGHFHDGHGFFFGDFHH?DFHDFFHHFHFHFF
```

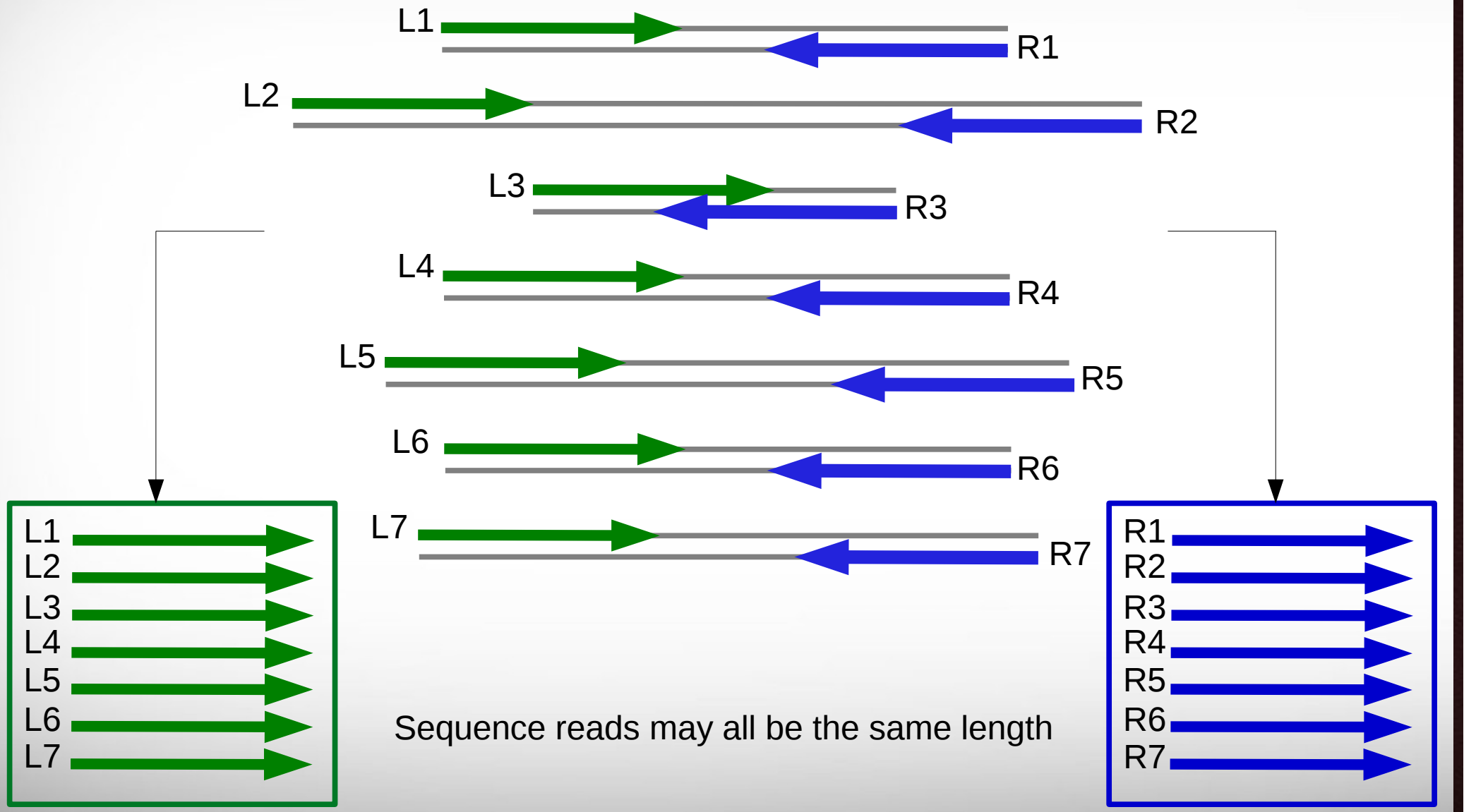
Read 2 pair fastq file

```
@M00861:1:000000000-A36BE:1:1101:14650:1529 2:N:0:8
ACTAAAAATCAATTTTATCAATTTCAAGCTCTACCTTATTTACTCATTATTTTAGTGATGGCCACTTTAATAAAAATATTGGTAGCATATTTTGCAATAG
+
BFFHIHHHFHHDGHIHHIHHHGHHHHHFHHDFFHIHIIIHIHDFHHHIHIIIH=AAFHIIIHFGFHHHHHGGHHIHHFGFFFEGGHHHDGHHH/CGHIFF
```

Variable Fragment Sizes



Maintain Read Pair Order

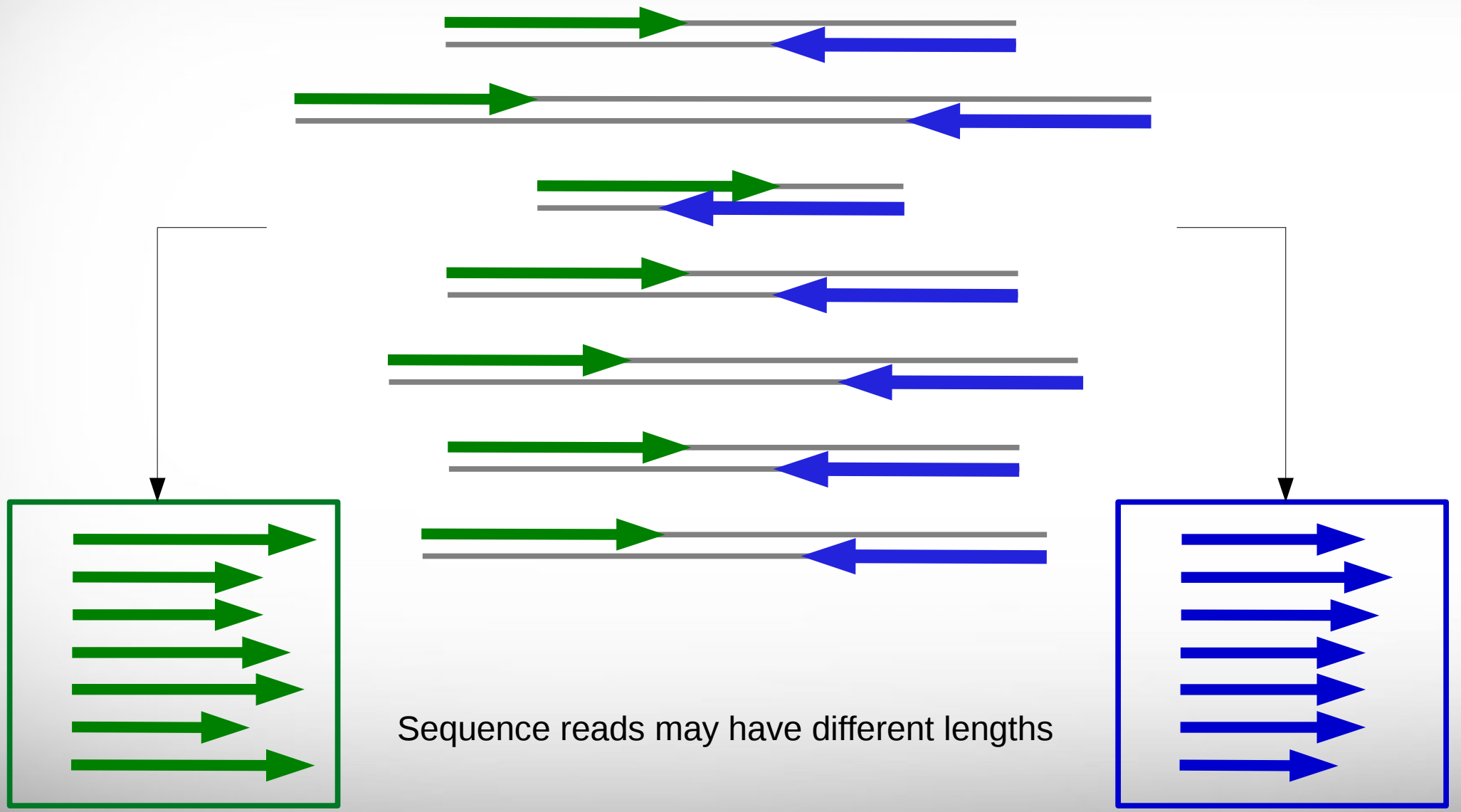


Sequence reads may all be the same length

Read 1 paired end fastq file

Read 2 paired end fastq file

MiSeq Can Perform Initial QC Trimming



Quality Control (QC)

QC Evaluation

- Use FastQC to visualize quality scores
  - Displays quality score distribution of reads
    - Input is a fastq file or files
    - Can disable grouping of sequence regions
  - Will alert you of poor read characteristics
  - Displays a representative sample of the fastq file
  - Can be run as a GUI or a command line interface

```
module spider fastqc
```

- FastQC will process using one CPU core per file
  - If there are 10 fastq files to analyze and 4 cores used
    - 4 files will start processing and 6 will wait in a queue
  - If there is only one fastq file to process then using 10 cores does not speed up the process

FastQC Exercise

- Use the GCATemplate for FastQC to submit a job evaluating the two sequence files

```
gedit run_fastqc_0.11.5_ada.sh &
```

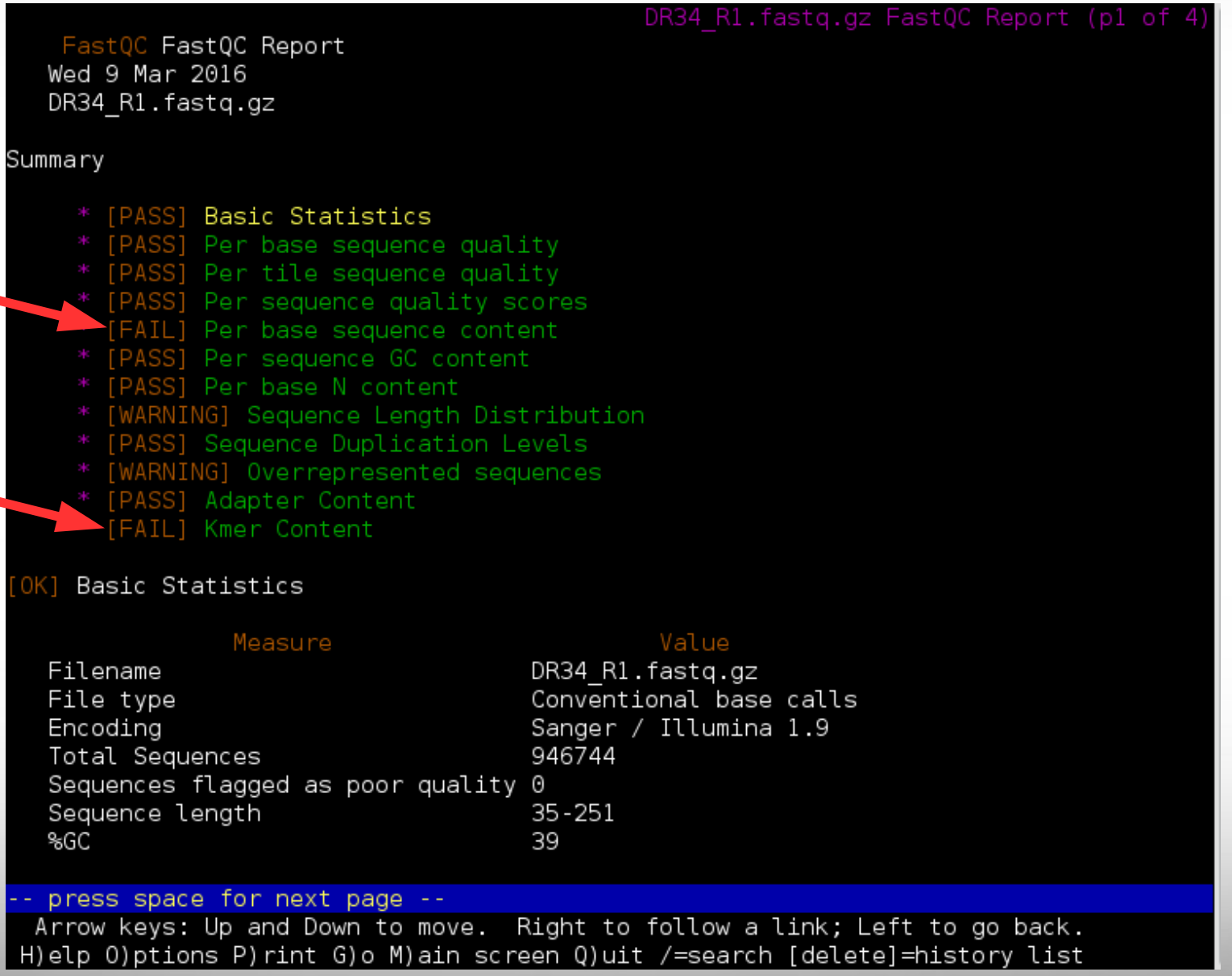
```
bsub < run_fastqc_0.11.5_ada.sh
```

- After your fastqc job is complete, unzip the results file and you can view the results files with **`lynx`** and **`eog`** (eog requires X11 login)

```
unzip DR34_R1_fastqc.zip
```

FastQC Report using lynx

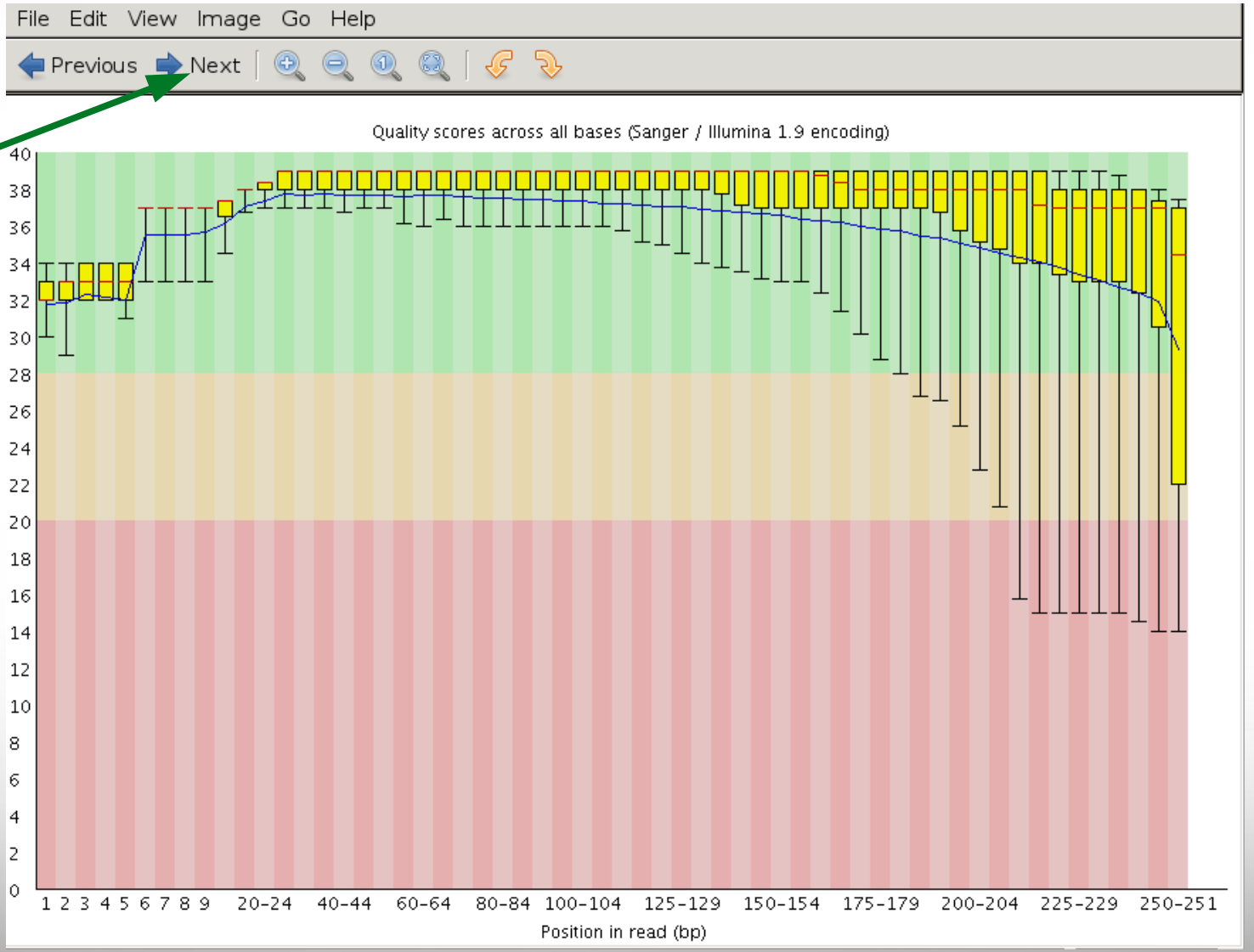
```
lynx DR34_R1_fastqc.html
```



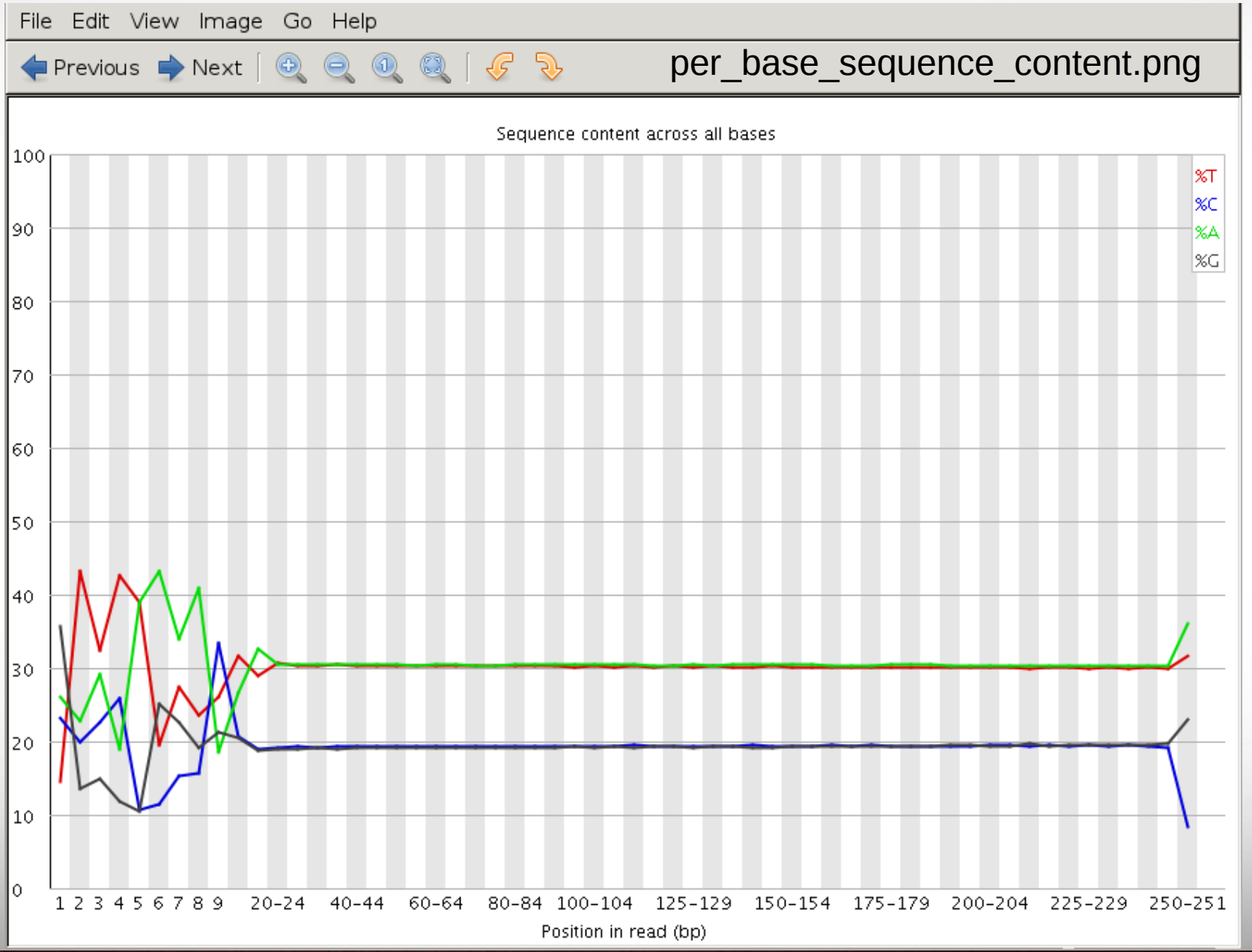
FastQC Output Image
Quality Distribution Before Trimming

```
eog DR34_R1_fastqc/Images/per_base_quality.png
```

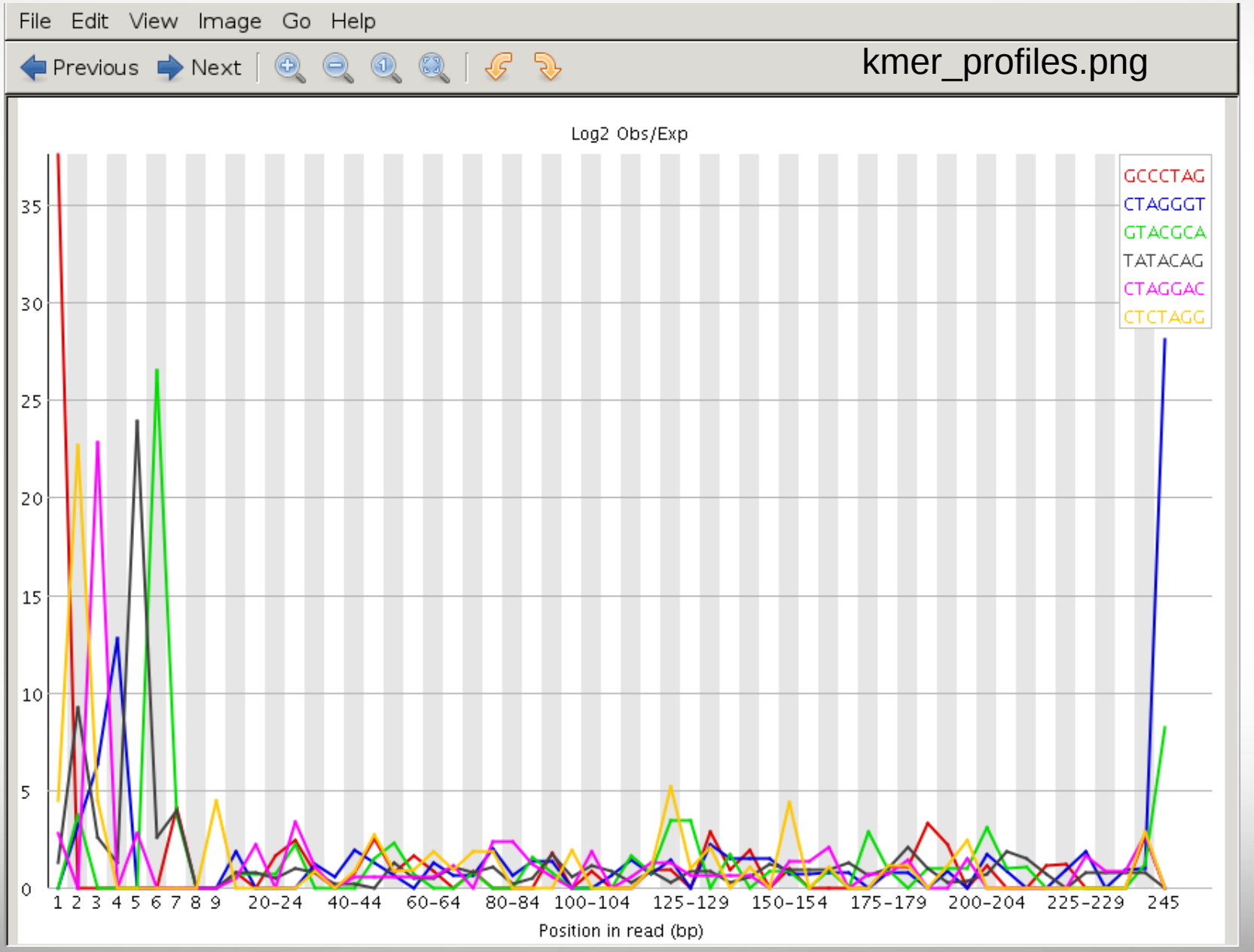
click for the next image in the same directory



Illumina Transposon Insertion Site

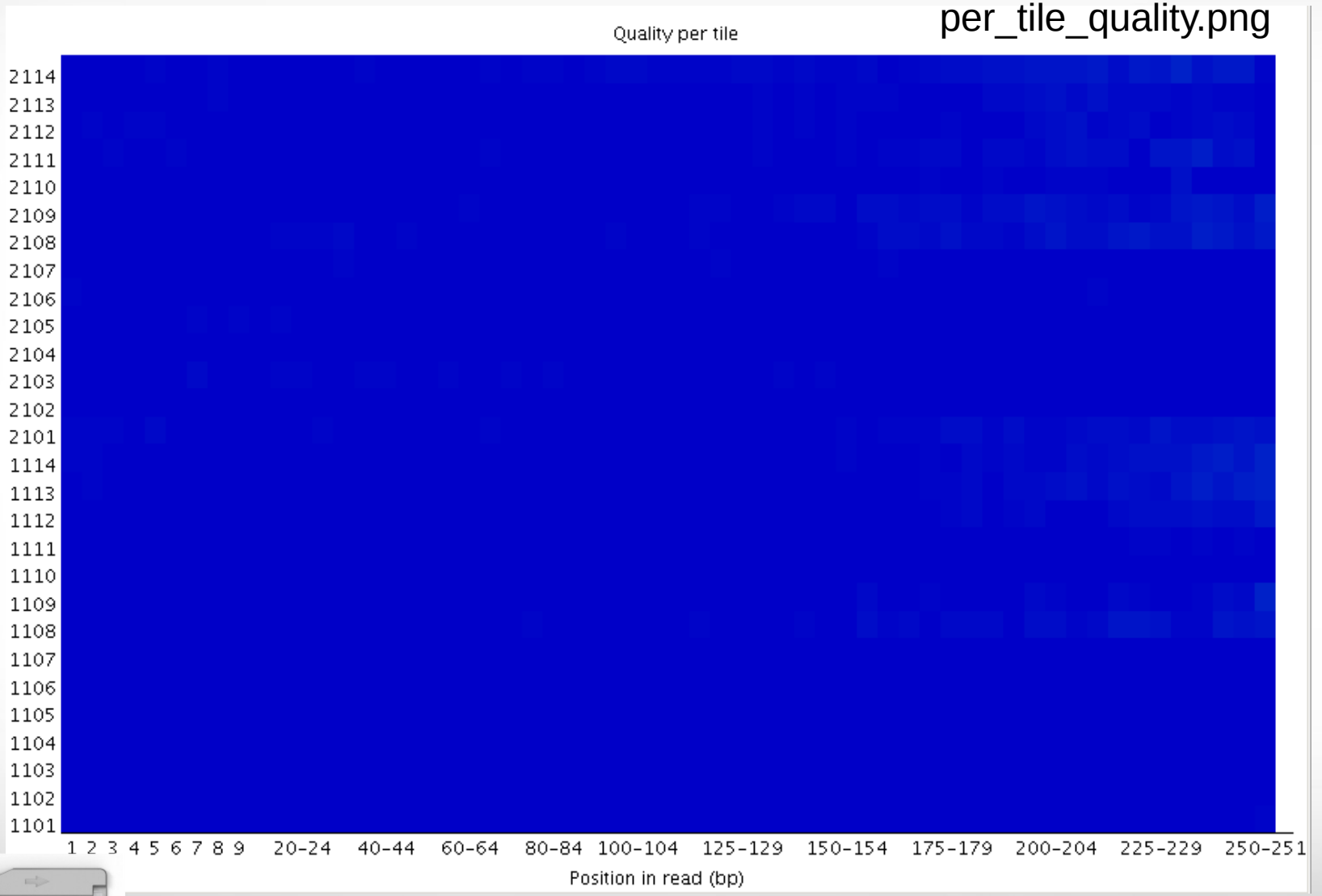


Illumina Transposon Insertion Site

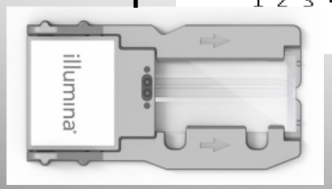


FastQC Output Image
Flowcell: good per_tile quality

per_tile_quality.png



MiSeq



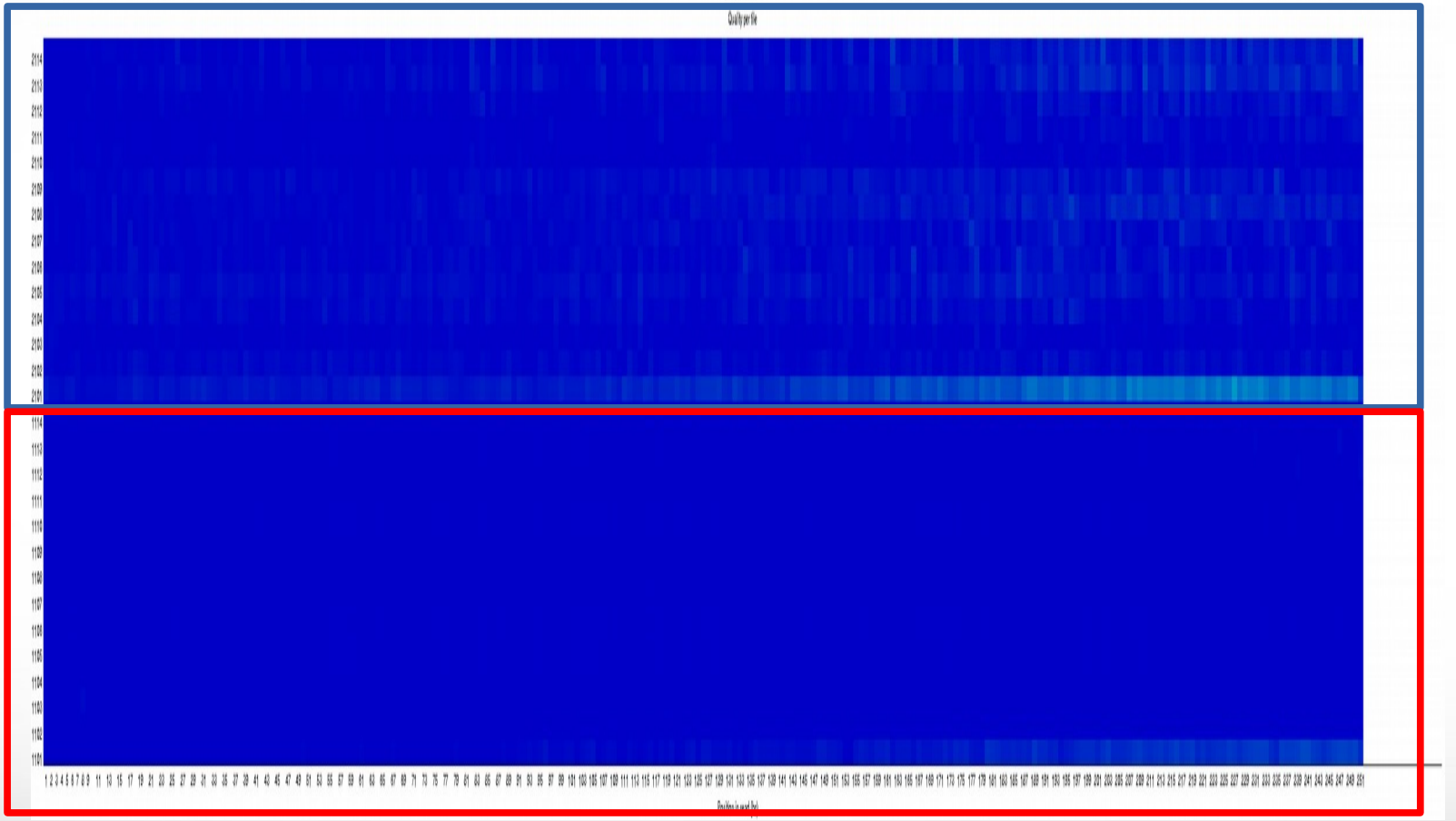
MiSeq flowcell

good quality  poor quality

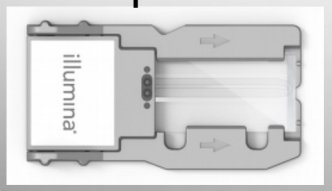
FastQC Output Image
Flowcell: good per_tile quality

bottom of flowcell

top of flowcell



MiSeq



MiSeq flowcell

good quality — poor quality

QC Quality Trimming

- Sequence quality trimming tools

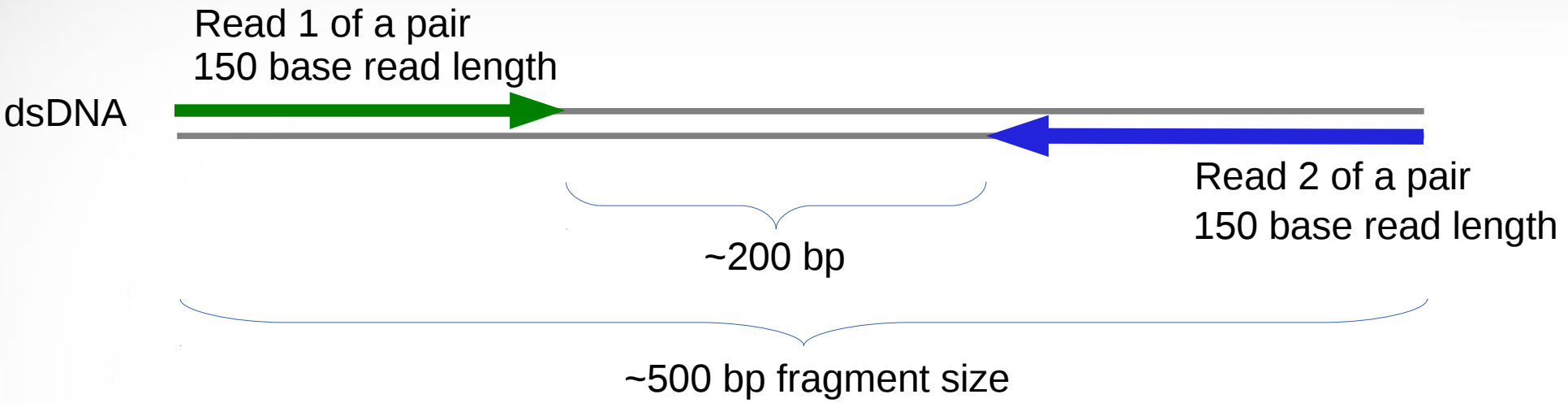
```
module spider Trimmomatic
```

← recommended tool

- Trimmomatic will maintain paired end read pairing after trimming
- Trim reads based on quality scores
  - Trim the same number of bases from each read or
  - Use a sliding window to calculate average quality at ends of sequences
- Decide if you want to discard reads with Ns
  - some assemblers replace Ns with As or a random base G, C, A or T
- Trim adapter sequences
  - Trimmomatic has a file of Illumina adapter sequences

```
ls $EBROOTTRIMMOMATIC/adapters
```

Paired End (PE) Reads



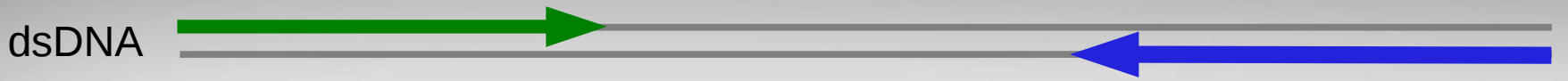
FASTQ format

Read 1 paired end fastq file

```
@M00861:1:000000000-A36BE:1:1101:14650:1529 1:N:0:8
TTCTTAAAAATACCATAAAAGGCTTAAACTTGCCATTTACGACGGATTAATTCCAACTCTTTTCGGCTATCTTCATCTTTTAAGGTTAAATGACTCATAA
+
FFFHBFFHHIIIIIIHFHHCGEFGHHIHHHIHD/?DGGHHH@DEB,5EGHGHHIIHIF?FGGHHCCBFDGHFHDGHGFFFGDFHH?DFHDFFHHFHFHFF
```

Read 2 paired end fastq file

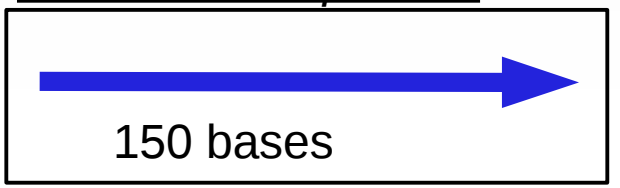
```
@M00861:1:000000000-A36BE:1:1101:14650:1529 2:N:0:8
ACTAAAAATCAATTTTATCAATTTCAAGCTCTACCTTATTTACTCATTATTTTAGTGATGGCCACTTTAATAAAAATATTGGTAGCATATTTTGCAATAG
+
BFFHIHHHFHHDGHIHHIHHHGHHHHHFHHDFFHIHIIIHIHDFHHHIHIIIH=AAFHIIIHFGFHHHHHGGHHIHHFGFFFEGGHHHDGHHH/CGHIFF
```



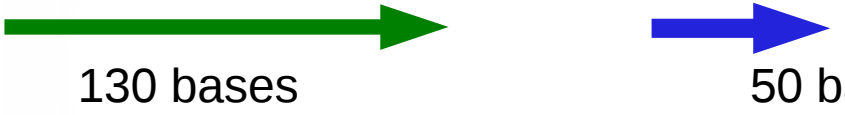
Trimming PE Sequence Reads

File 1 from sequencer

File 2 from sequencer



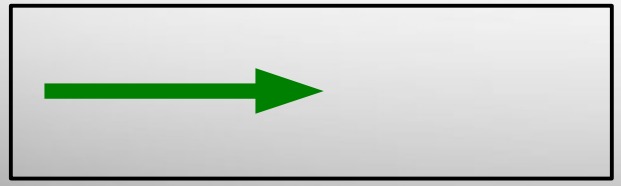
Length after trimming



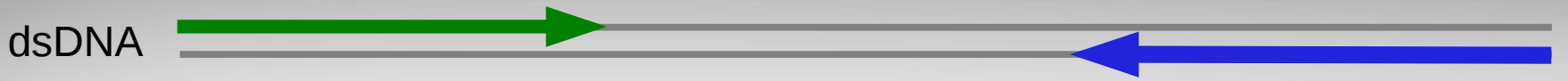
minimun read length = 40

FASTQ Files with trimmed reads

Paired end 1 file



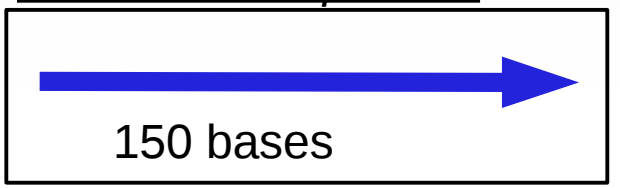
Paired end 2 file



Trimming PE Sequence Reads

File 1 from sequencer

File 2 from sequencer



Length after trimming

minimun read length = 40

FASTQ Files with trimmed reads

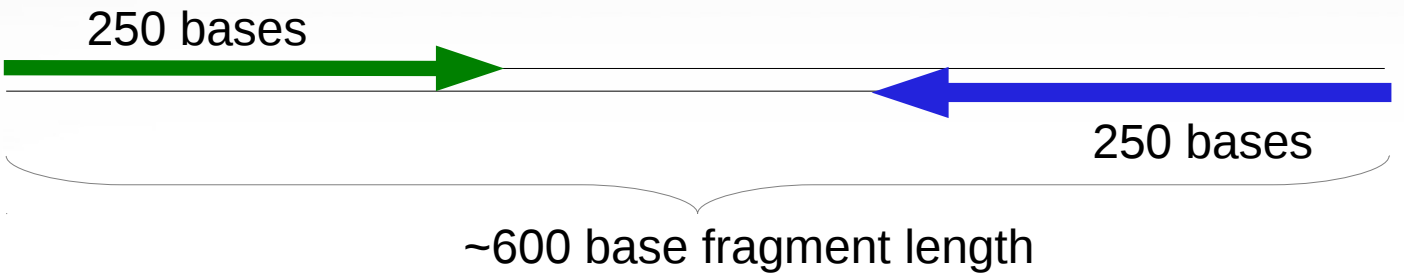
Paired end 1 file

Paired end 2 file

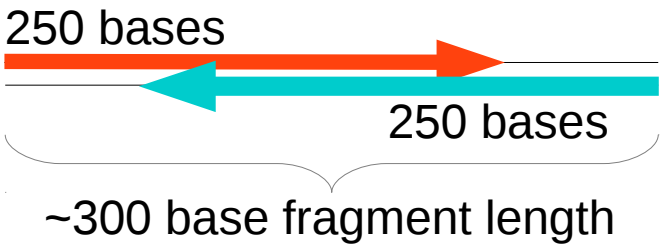
Single end file (unpaired)

Merge Overlapping Paired End Reads

fragment 1
dsDNA

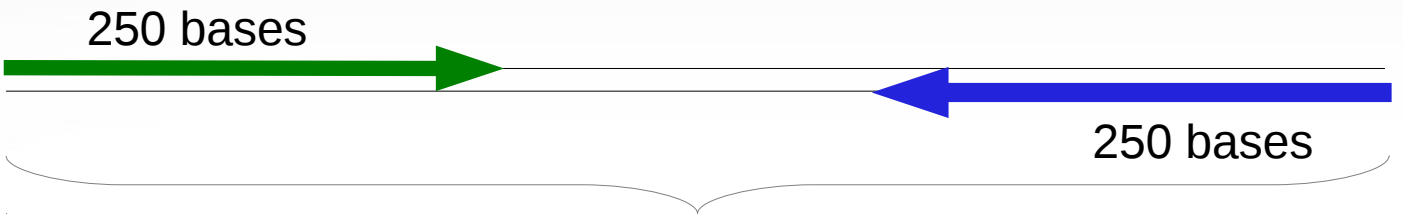


fragment 2
dsDNA

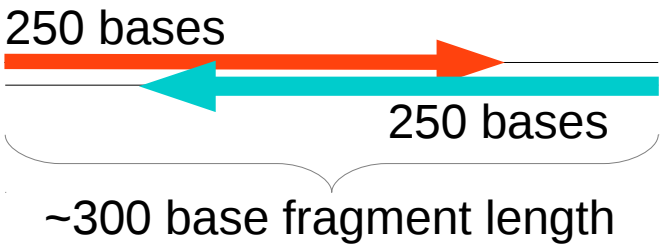


Merge Overlapping Paired End Reads

fragment 1
dsDNA



fragment 2
dsDNA



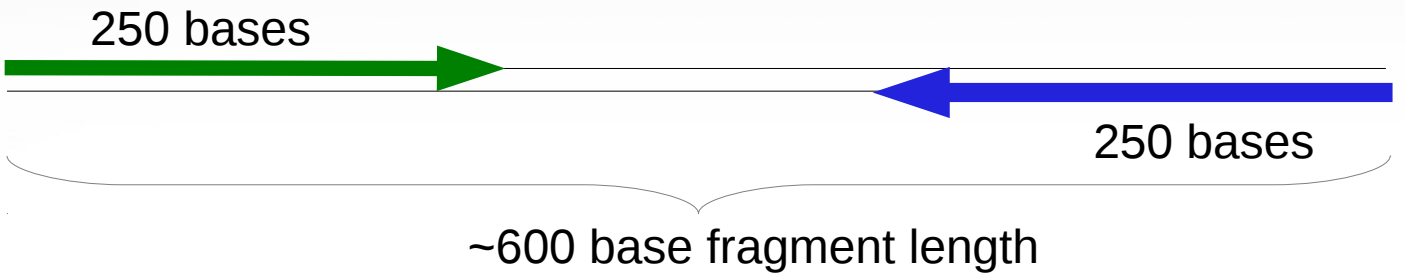
FASTQ Files with trimmed reads

Pair 1 file

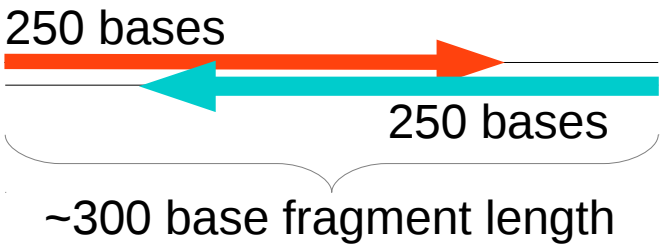
Pair 2 file

Merge Overlapping Paired End Reads

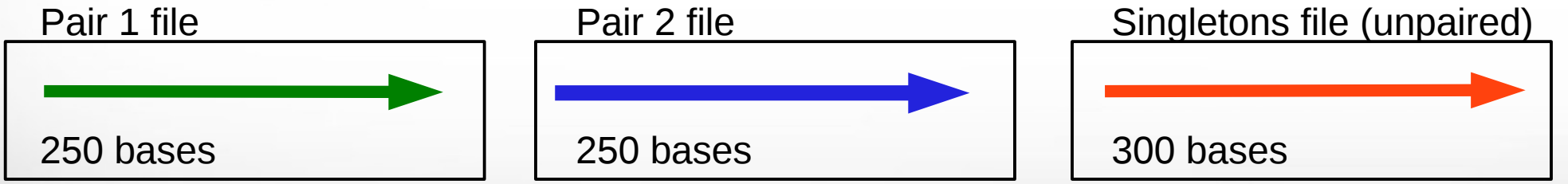
fragment 1
dsDNA



fragment 2
dsDNA



FASTQ Files with trimmed reads



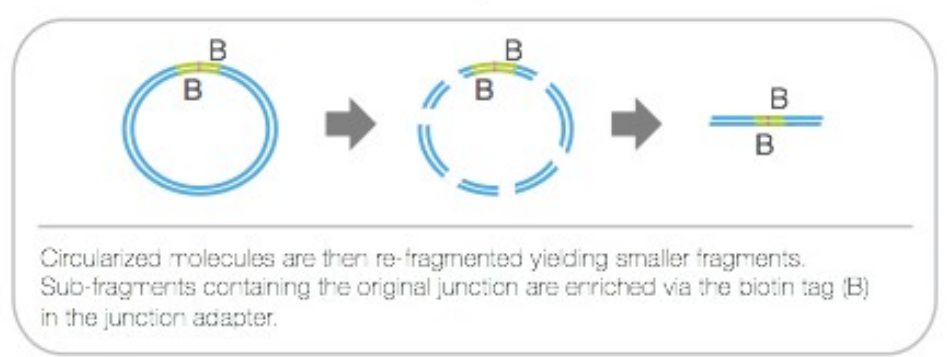
Tools for merging overlapping reads:

```
module spider FLASH
module spider Coperead
module spider PEAR
```

Verifying and Trimming Mate Pair Reads

Biotin junction

adapter sequence 38 bp



illumina.com

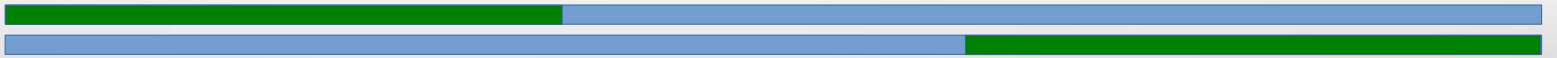
true mate pair

true mate pair

true mate pair

?

?



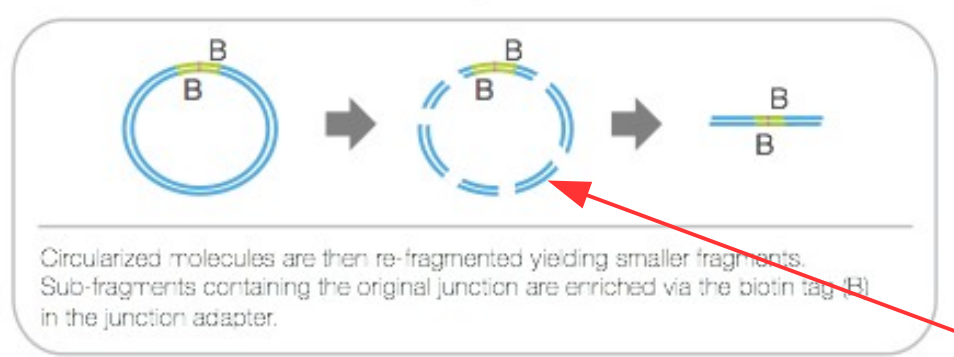
(paired end contaminant)

Tool for trimming Illumina Nextera mate pairs: `module spider NextClip`

Verifying and Trimming Mate Pair Reads

Biotin junction

adapter sequence 38 bp



illumina.com

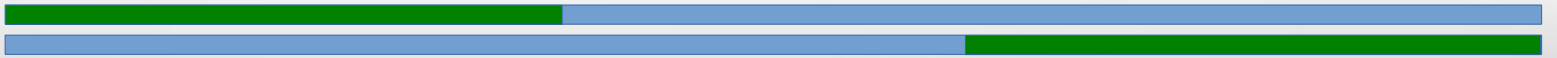
true mate pair

true mate pair

true mate pair

?

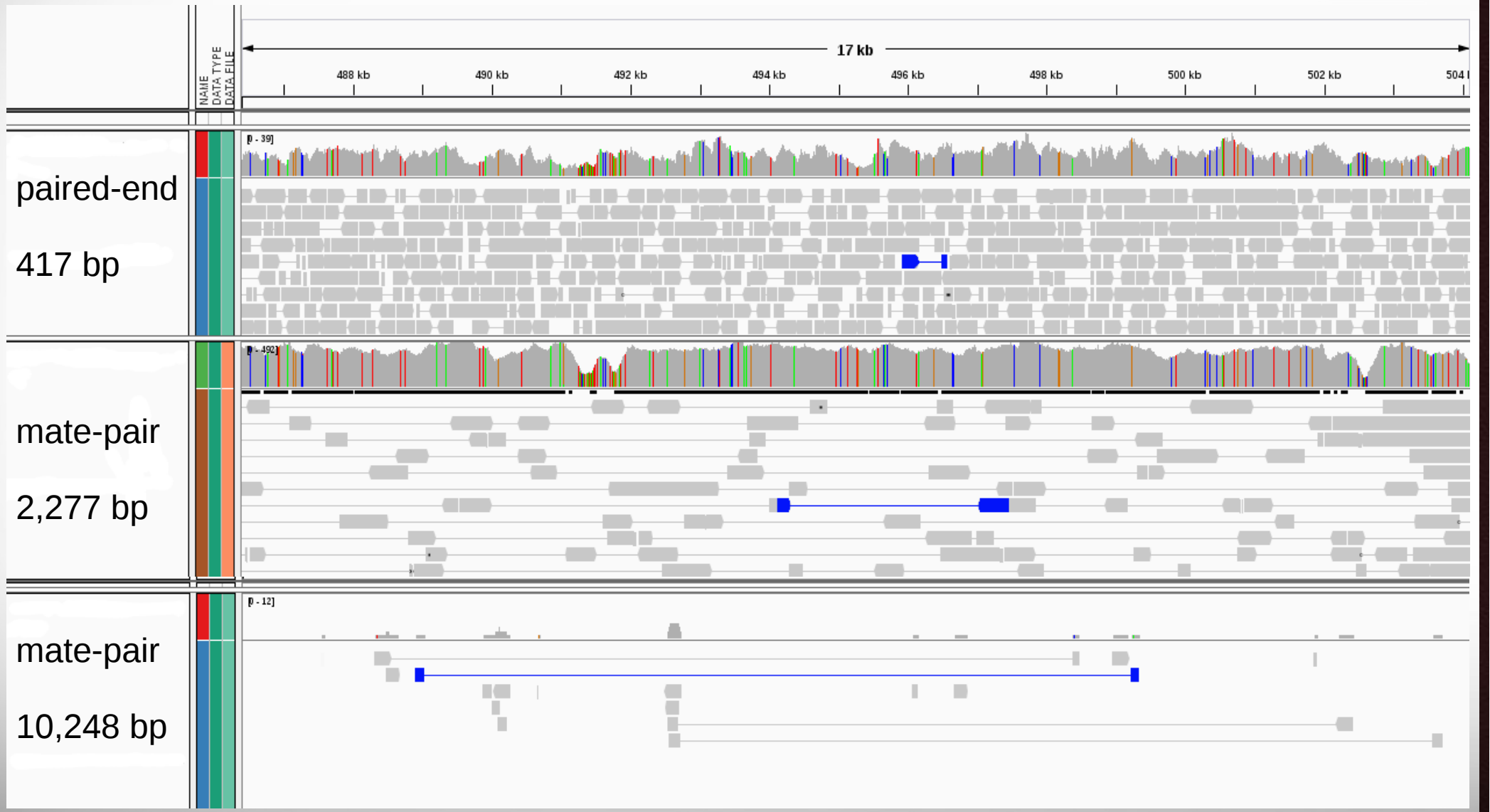
?



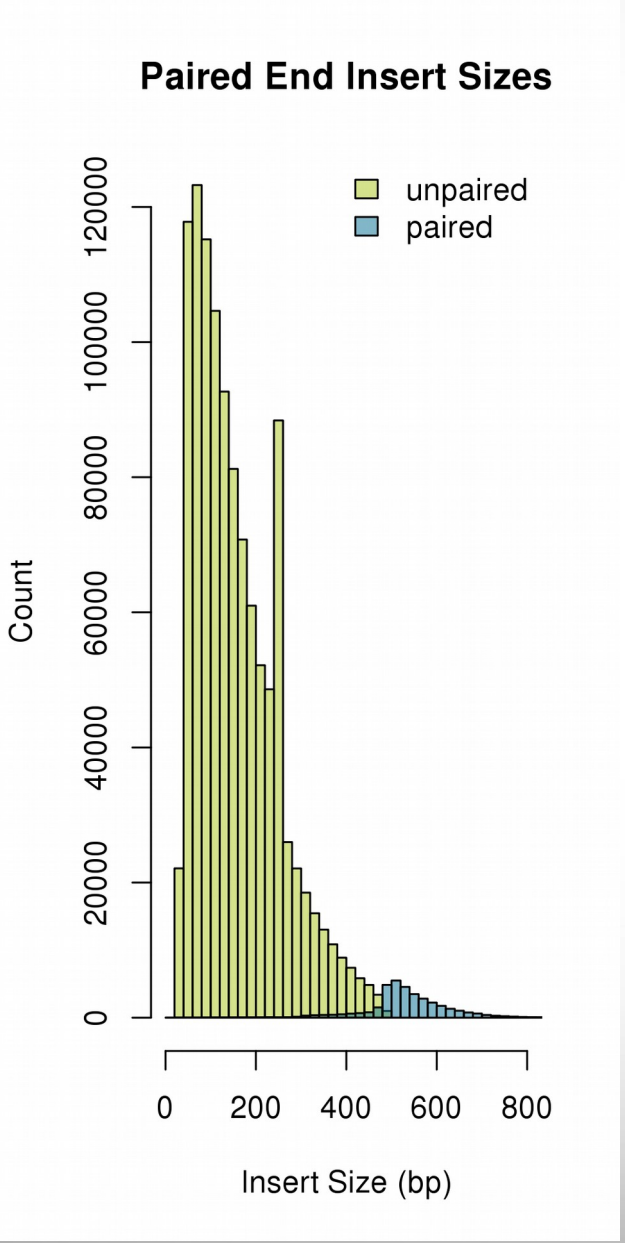
(paired end contaminant)

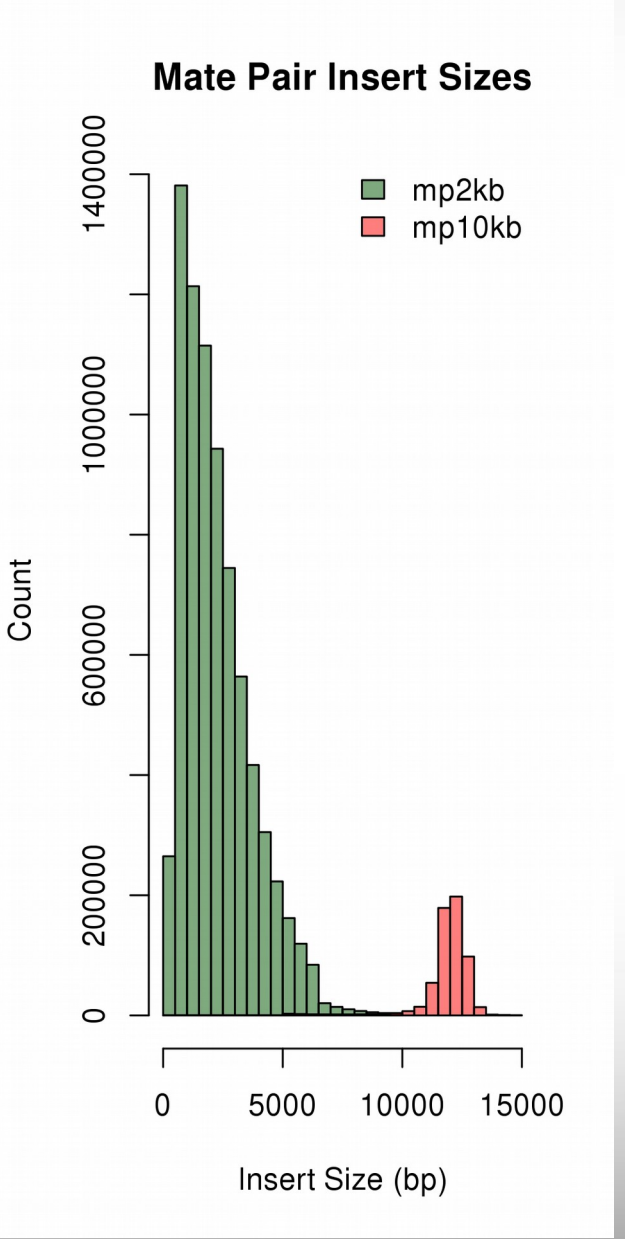
Tool for trimming Illumina Nextera mate pairs: `module spider NextClip`

Sequencing Libraries Insert Sizes



Sequence Library Insert Size Distribution

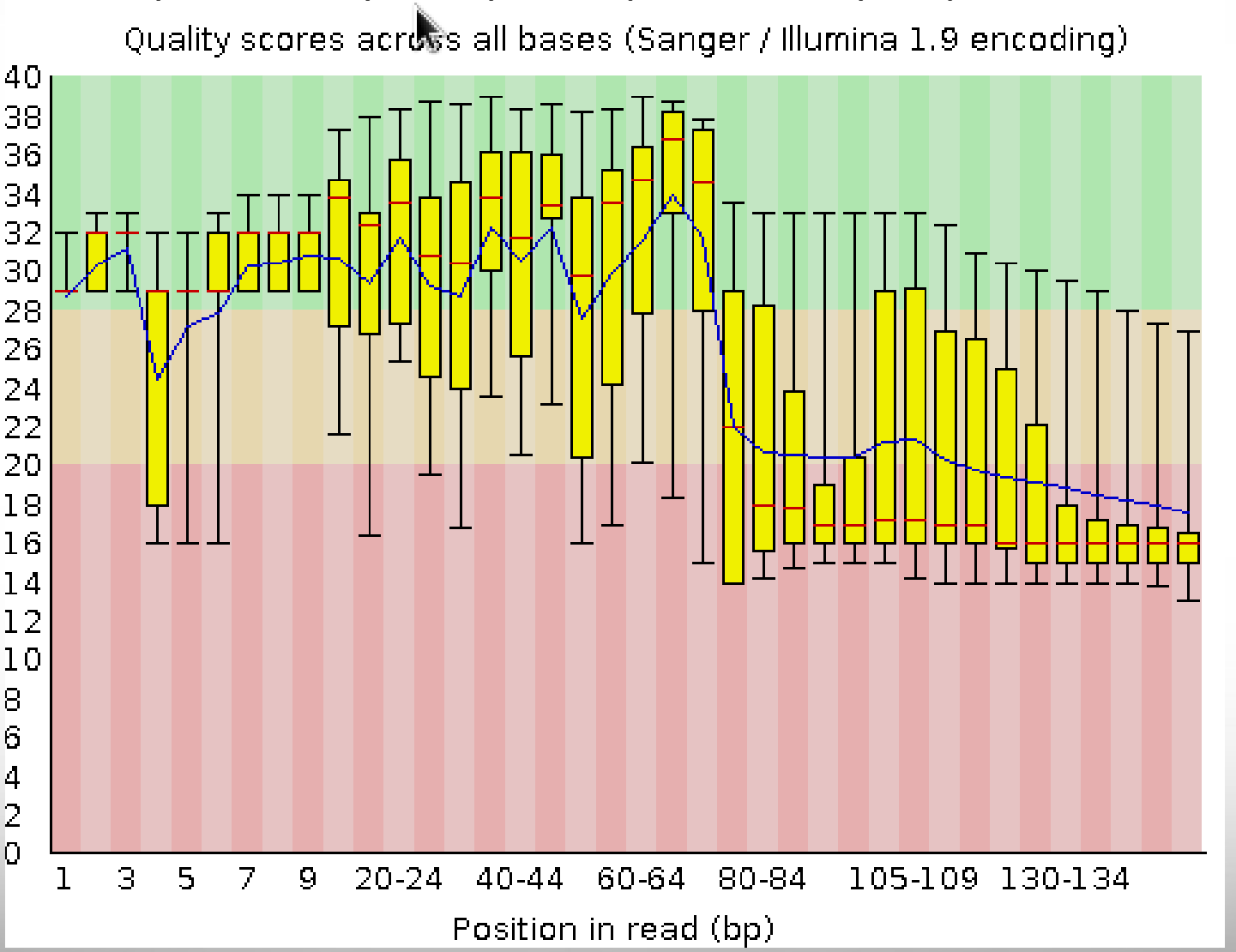




Failed QC Examples

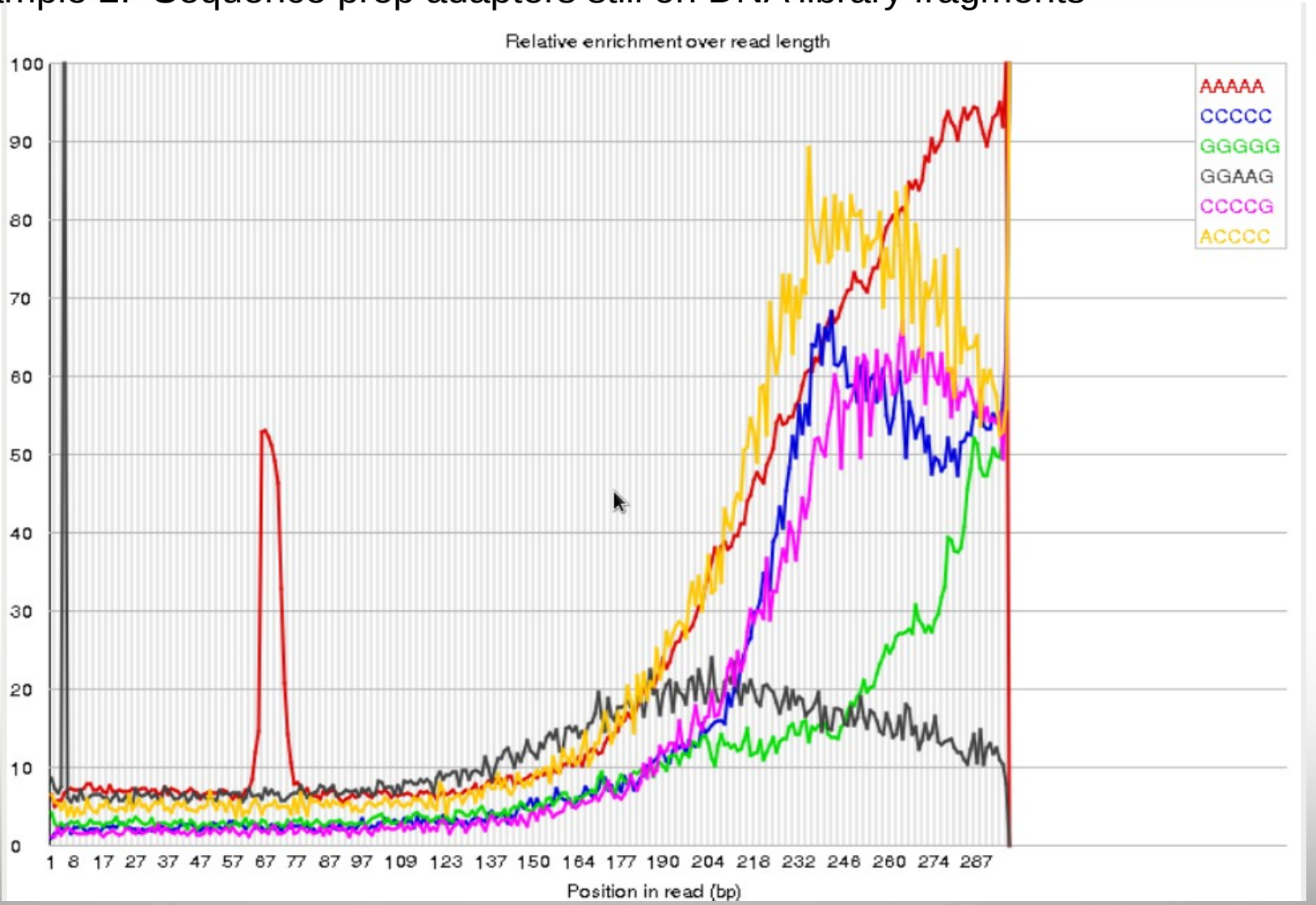
FastQC Output Image
Failed Per base sequence quality

Example 1. Expired MiSeq mate pair kit (9 months expired)



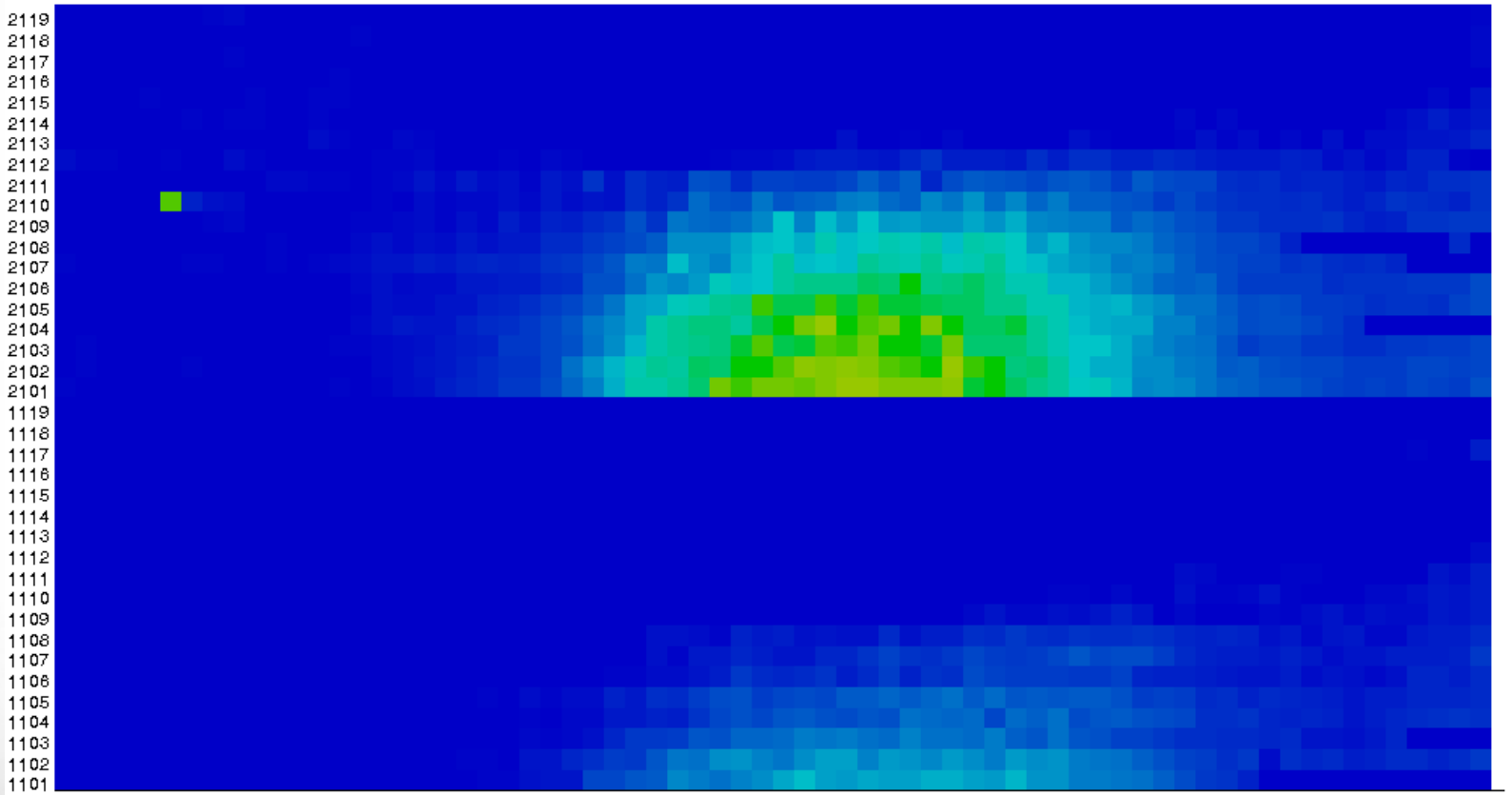
FastQC Output Image
Failed Kmer Content

Example 2. Sequence prep adapters still on DNA library fragments

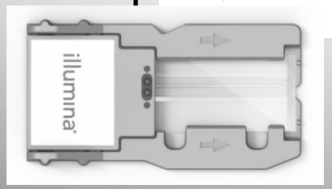


FastQC Output Image
Flowcell: not good per_tile quality

Example 3. Faulty flowcell



MiSeq



good quality — poor quality

MiSeq flowcell

Mapping Reads to a Reference Assembly

Mapping Reads to a Reference Assembly

- Align reads using bwa
  - `module spider BWA`
    - bwa index files for UCSC genomes found here
    - `/scratch/datasets/genome_indexes/ucsc/mm10/bwa_0.7.12_index/`
- Align reads using bowtie or bowtie2
  - `module spider Bowtie`
    - Bowtie index files for UCSC genomes found here:
    - `/scratch/datasets/genome_indexes/ucsc/mm10/bowtie_index/`
  - `module spider Bowtie2`
    - Bowtie2 index files for UCSC found here:
    - `/scratch/datasets/genome_indexes/ucsc/mm10/bowtie2_index/`

Visualize bam File Alignments

Sorting, Viewing sam/bam Files

- Sequence Alignment/Map format (sam)
  - view sam files using the UNIX command:

```
more file.sam
```

- Binary Alignment/Map format (bam)
  - Compressed (binary) sam files need samtools to view

```
module load SAMtools/1.3-intel-2015B
```

  - Recommended: sort sam/bam file based on coordinate into bam format

```
samtools sort -@ 10 -m 1G -o file_sorted.bam file.sam
```

  - Create an index of the bam file using samtools

```
samtools index file_sorted.bam
```

    - A samtools index is needed prior to viewing bam files in browsers
  - Viewing bam files using samtools

`samtools view file_sorted.bam | more`	view only alignments
`samtools view -H file_sorted.bam`	view only header
`samtools view -h file_sorted.bam | more`	view header + alignments

Alignment Statistics

```
samtools flagstat file_sorted.bam
```

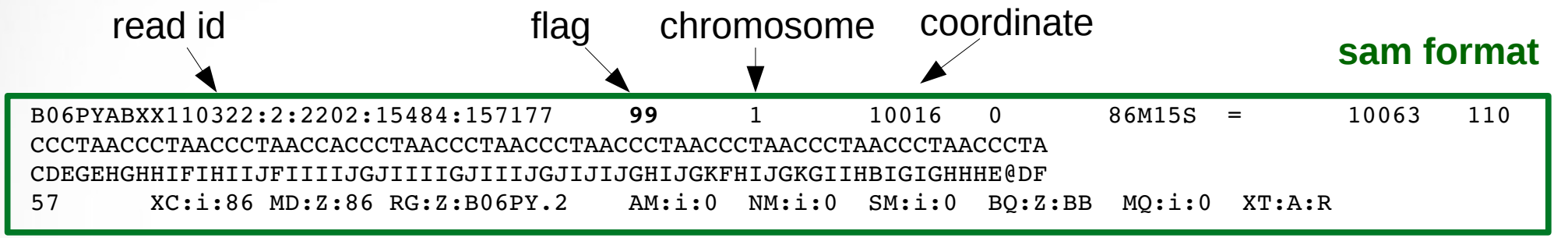
```
2152003688 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
109088892 + 0 duplicates
2125295307 + 0 mapped (98.76% : N/A)
2152003688 + 0 paired in sequencing
1076001844 + 0 read1
1076001844 + 0 read2
2077082056 + 0 properly paired (96.52% : N/A)
2098587704 + 0 with itself and mate mapped
26707603 + 0 singletons (1.24% : N/A)
11058620 + 0 with mate mapped to a different chr
7628294 + 0 with mate mapped to a different chr (mapQ>=5)
```

Both reads in the pair are mapped on the same chromosome and in FR or RF orientation

Sam Flags

https://broadinstitute.github.io/picard/explain-flags.html

- Flags describe alignment (flag is the sum of bits)



bits: **1** **2** 4 8 16 **32** **64** 128 256 512 1024 2048

99 = 64 + 32 + 2 + 1

- Filter bam alignments based on bit in flag (-f and/or -F)
  - Keep only reads that are 'mapped in proper pair'

```
samtools view -h -b -f 2 file.bam > paired_reads.bam
```

  - Keep all except reads that are 'PCR or optical duplicate'

```
samtools view -h -b -F 1024 file.bam > dedup_reads.bam
```

Sample bam and reference files

For this samtools demo, add symbolic links* to the example files in your working directory

```
ln -s /scratch/helpdesk/ngs/alignments/cdubl/dr34_sorted.bam
```

Add a symbolic link to the example reference genome fasta file

```
ln -s /scratch/helpdesk/ngs/genomes/candida/c_dubliniensis.fa
```

Use the tab key when typing these long paths

* The symbolic links are used to make the commands shorter for demonstration purposes only. You do not need to make symbolic links in order to use **`samtools tview`**

SAMtools without a Reference Genome

Reference genome displayed on top as NNNNNNNNNN

```
module load SAMtools/1.3-intel-2015B
```

```
samtools index dr34_sorted.bam
```

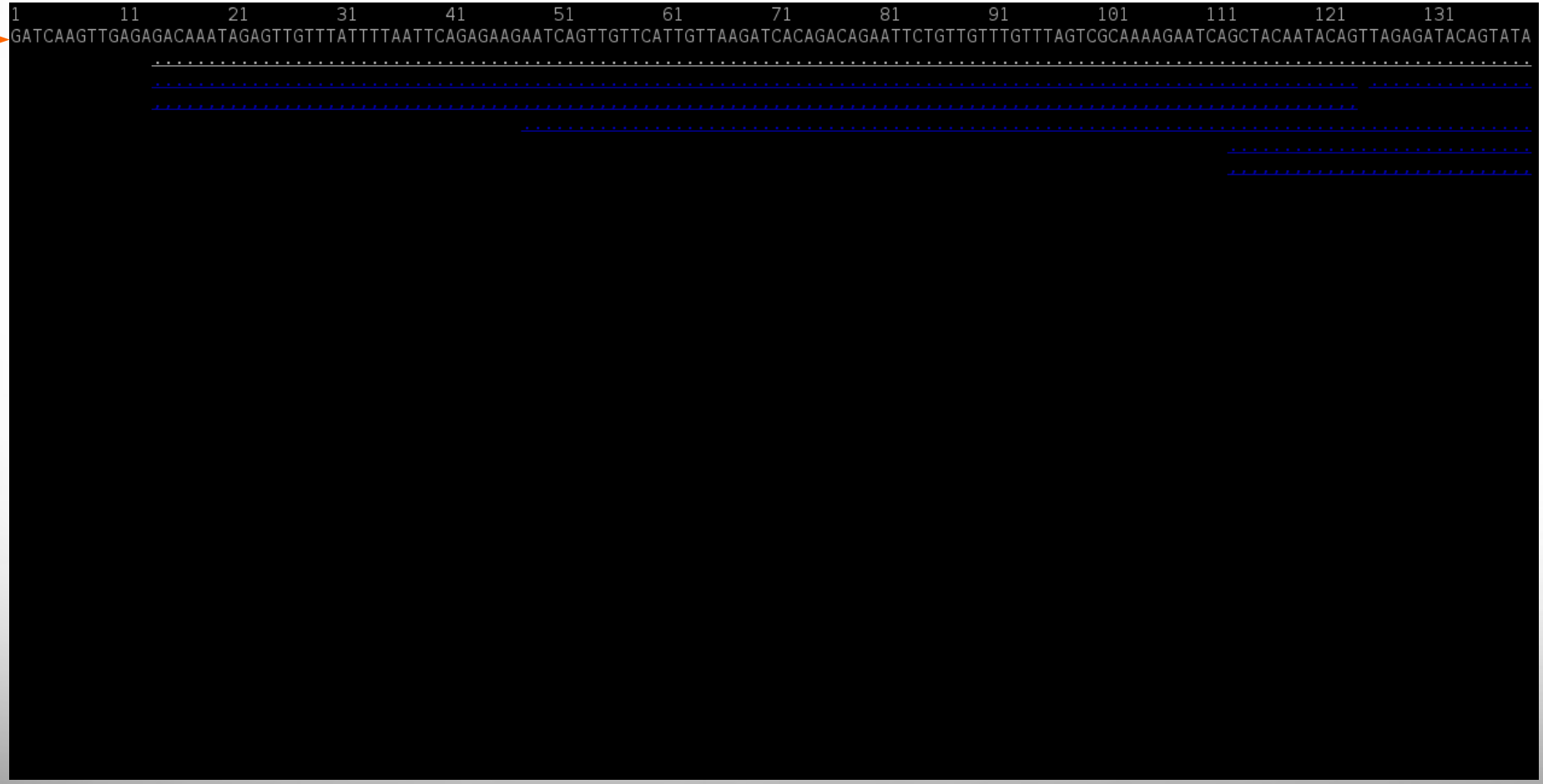
```
samtools tview dr34_sorted.bam
```

```
1         11        21        31        41        51        61        71        81        91        101       111       121       131
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
             GACAAATAGAGTTGTTTATTTTAATTCAGAGAAGAATCAGTTGTTCATTGTTAAGATCACAGACAGAATTCTGTTGTTTGTTTAGTCGCAAAAGAATCAGCTACAATACAGTTAGAGATACAGTATA
             GACAAATAGAGTTGTTTATTTTAATTCAGAGAAGAATCAGTTGTTCATTGTTAAGATCACAGACAGAATTCTGTTGTTTGTTTAGTCGCAAAAGAATCAGCTACAATACAG TAGAGATACAGTATA
             gacaaatagagttgtttattttaattcagagaagaatcagttgttcattgttaagatcacagacagaattctgttgtttgtttagtcgcaaaagaatcagctacaatacag
                                               AATCAGTTGTTCATTGTTAAGATCACAGACAGAATTCTGTTGTTTGTTTAGTCGCAAAAGAATCAGCTACAATACAGTTAGAGATACAGTATA
                                                                                                                GCTACAATACAGTTAGAGATACAGTATA
                                                                                                                gctacaatacagttagagatacagtata
```

SAMtools with a Reference Genome

Reference genome sequence displayed on top

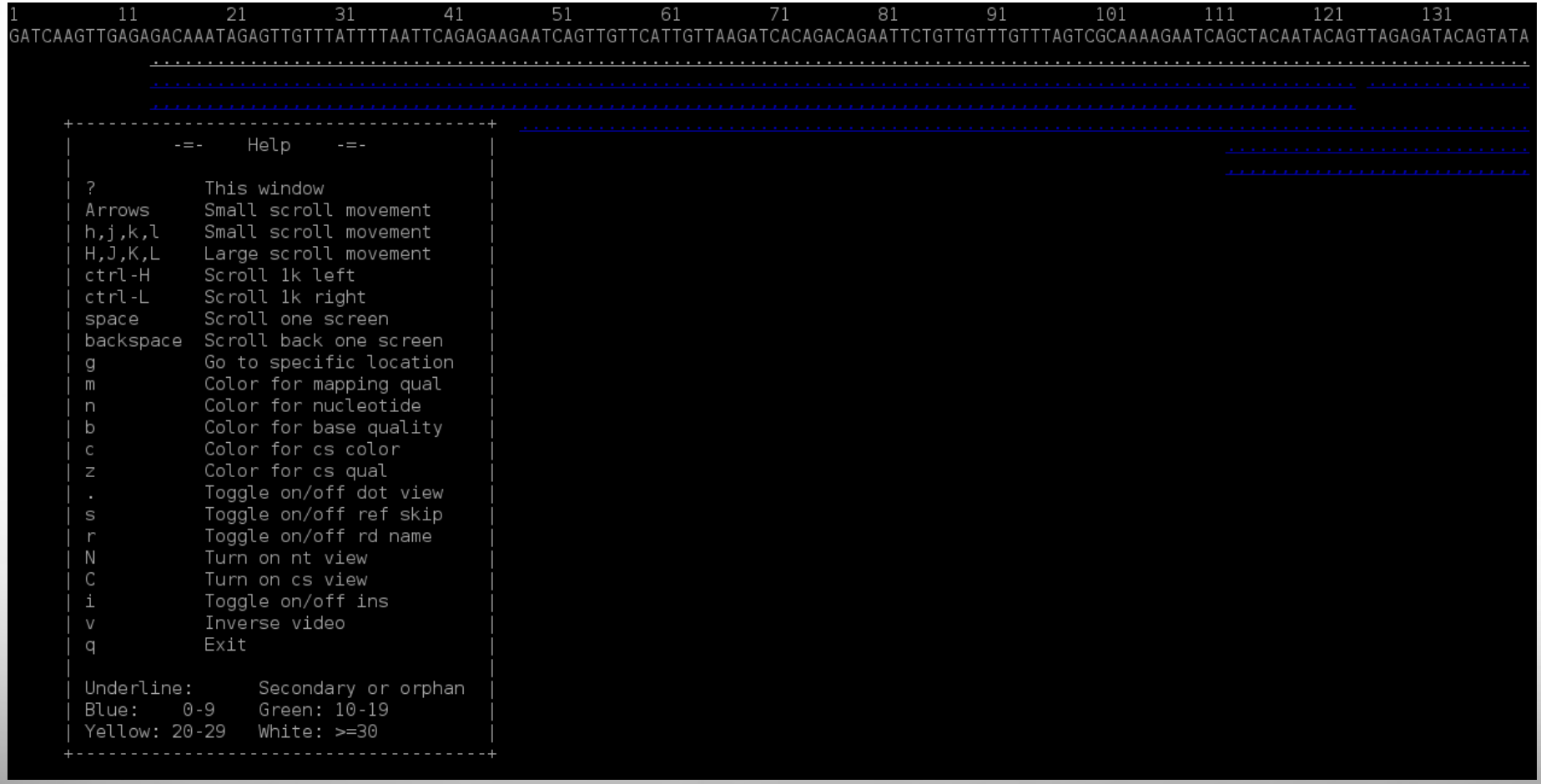
```
samtools tview dr34_sorted.bam c_dubliniensis.fa
```



SAMtools with a Reference Genome

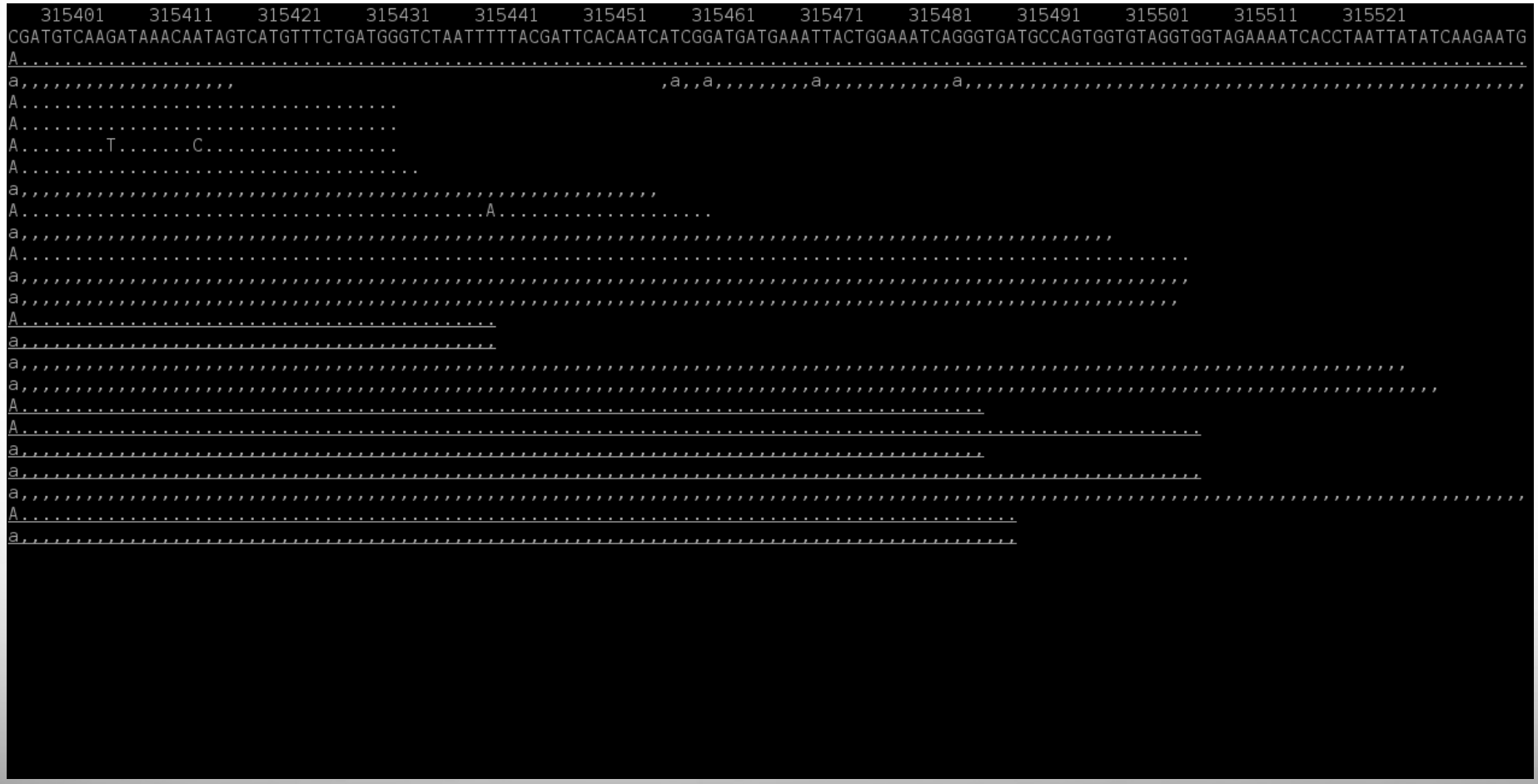
Type ? for help menu

```
samtools tview dr34_sorted.bam c_dubliniensis.fa
```



View at a Specific Coordinate

```
samtools tview dr34_sorted.bam c_dubliniensis.fa -p 1:315398
```



Integrative Genomics Viewer (IGV) Exercise

IGV is a genome browser with pre-loaded genomes available in which you can use to view multiple .bed, .sam and .vcf files.
IGV is launched from a login node not a job script or compute node.

```
module spider IGV
```

```
module load IGV/2.3.68-Java-1.8.0_66
```

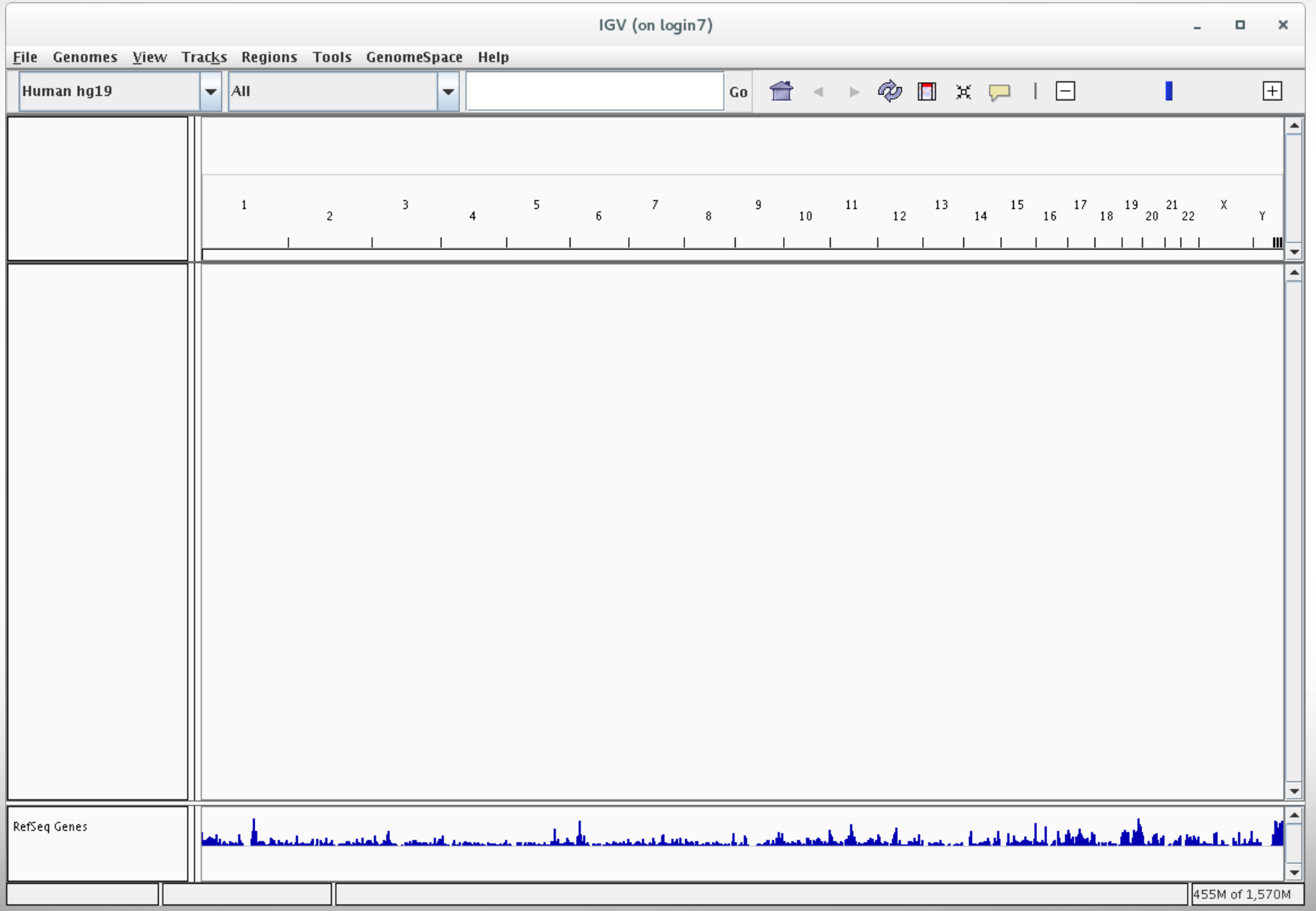
Launch IGV using the igv.sh script (X11 login required)

```
igv.sh
```

bam files need to be indexed prior to viewing with IGV.
Index bam files with samtools which will create an index file that will have the extension **.bam.bai**

```
module load SAMtools/1.3-intel-2015B
samtools index my_file.bam
```

hg19 is default Reference Genome

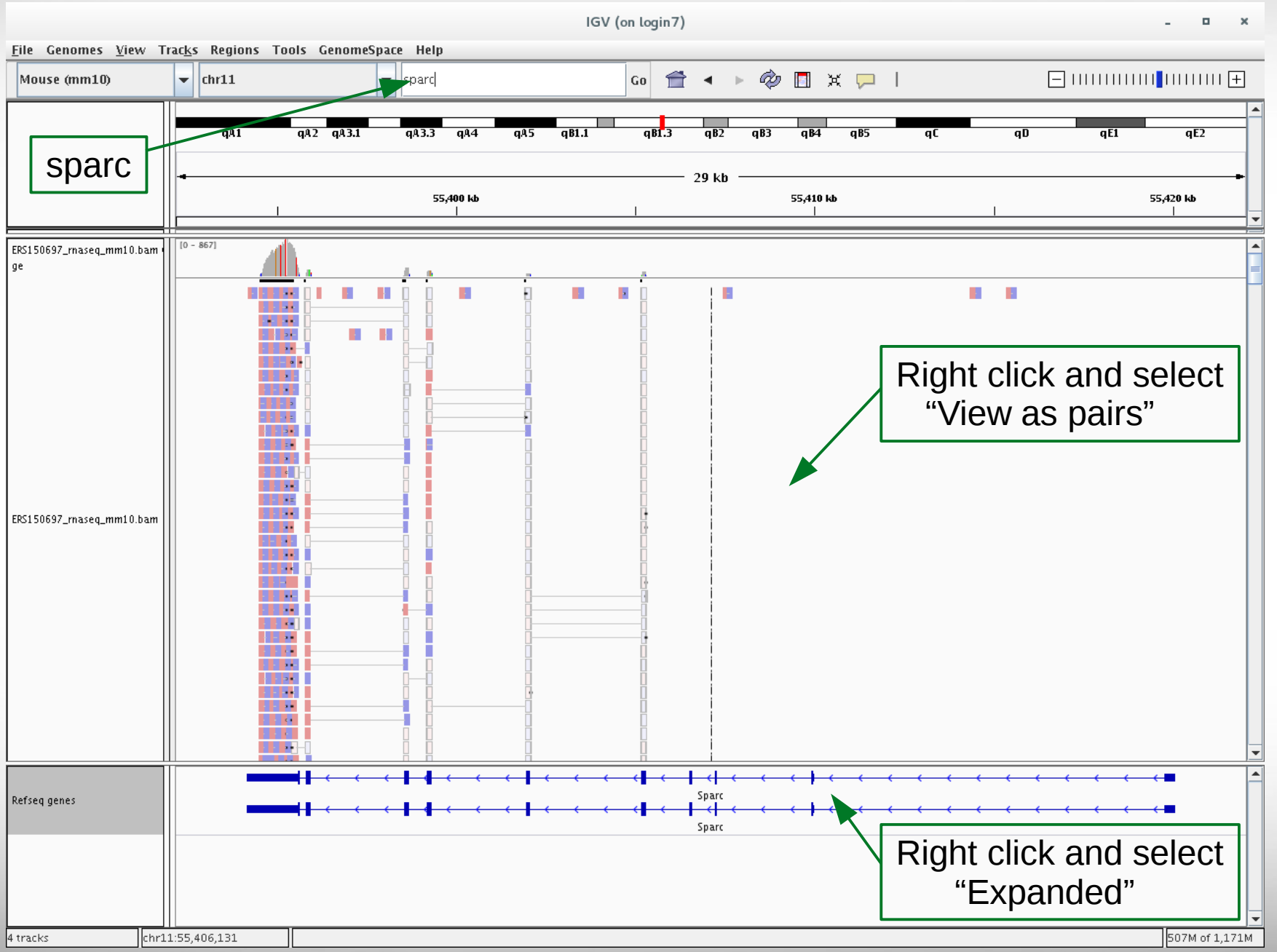


Change the Reference Genome

Load bam Alignment File

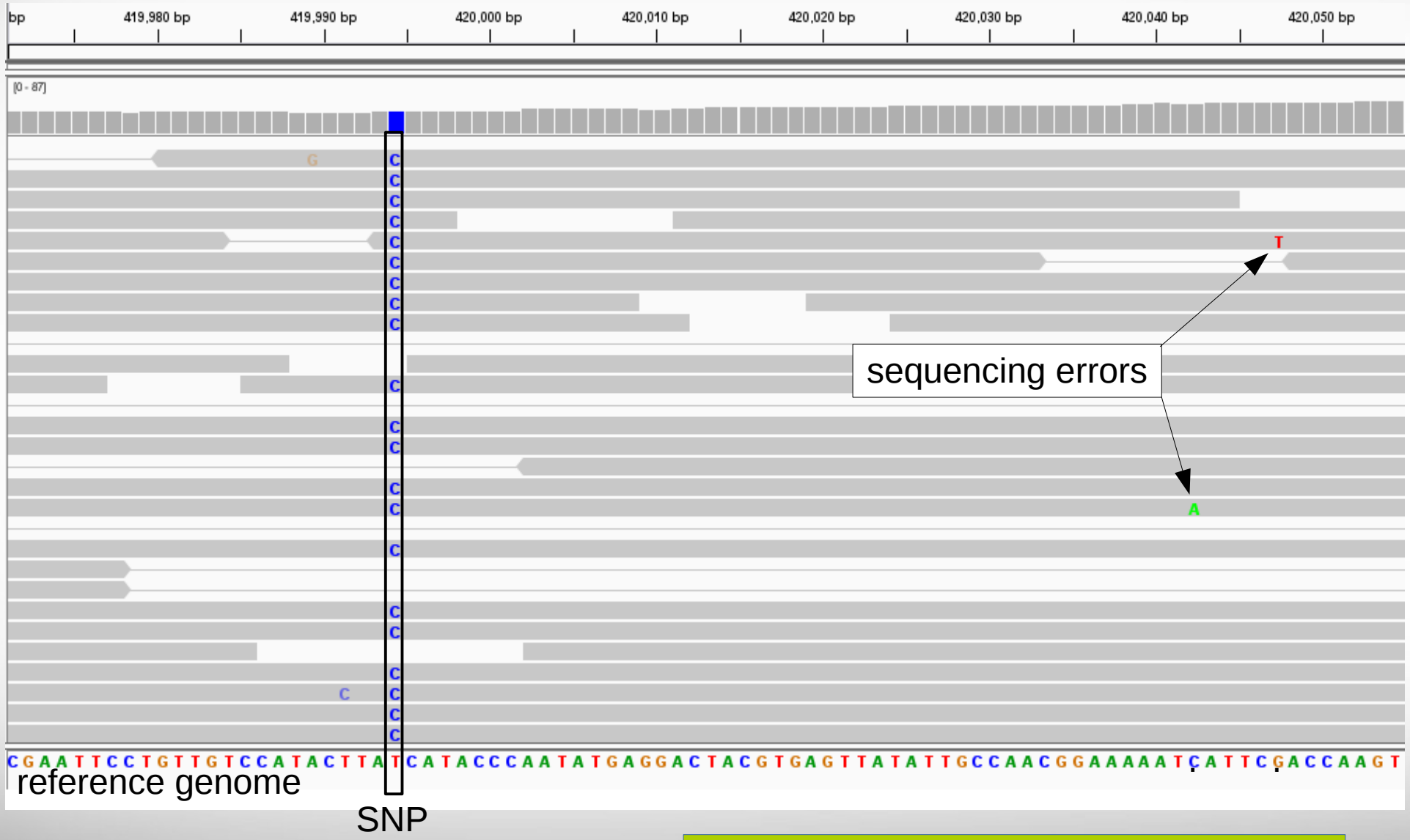
- Select “File → “Load from file”
- Search for
  - /scratch/helpdesk/ngs/alignments/mm10/ERS150697_rnaseq_mm10.bam

IGV viewing indexed bam file



Sequence Error Correction

Sequencing Errors



Tool for correcting sequencing errors

```
module spider Lighter
```

Digital Normalization

Digital Normalization

- Reduce memory requirements by reducing the number of redundant sequence reads if you have a very high sequencing coverage (> 200x)

```
module spider BBMap
```

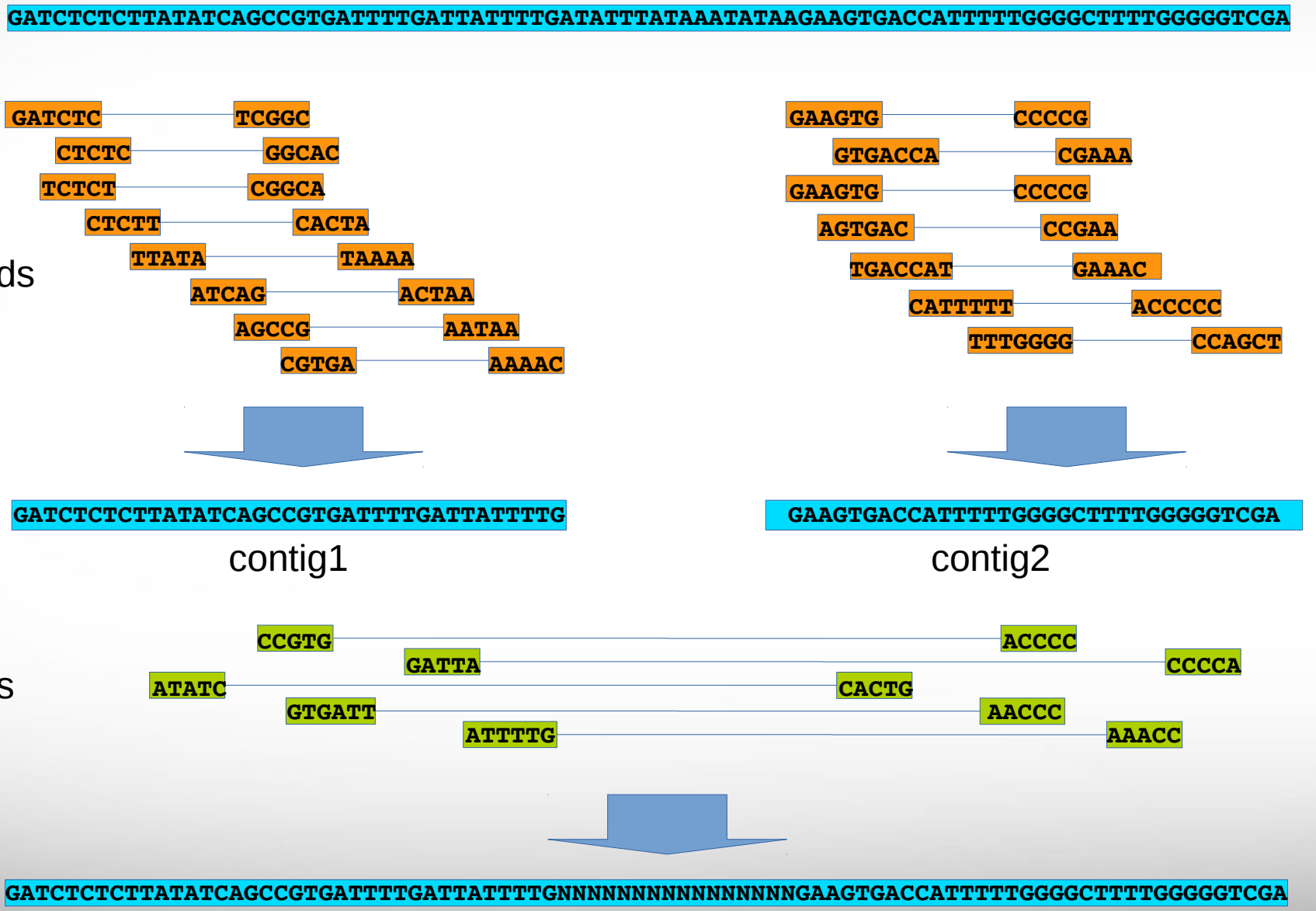
Use the bbnorm.sh script in the BBMap module

A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data

C. Titus Brown[1,2,*], Adina Howe[2], Qingpeng Zhang[1], Alexis B. Pyrkosz[3], Timothy H. Brom[1]
1 Computer Science and Engineering, Michigan State University, East Lansing, MI, USA
2 Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA
3 USDA Avian Disease and Oncology Laboratory, East Lansing, MI, USA
* E-mail: ctb@msu.edu

de novo Genome Assembly

Assembly Results in Contigs and Scaffolds



Genome Assemblers Assemble using k-mers

If you set kmer size = 31 then reads shorter than 31 bp will not be used in the assembly

de novo Assemblers Installed on Ada

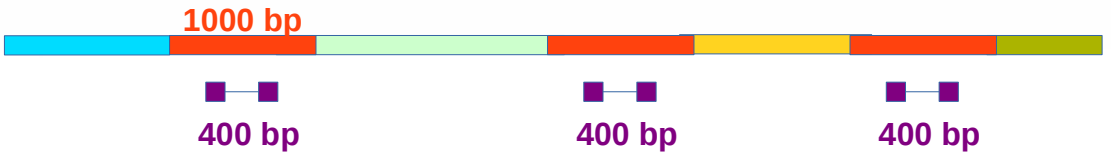
- `module spider SPAdes` small genomes
- `module spider MaSuRCA` any size genome
- `module spider ABySS` any size genome
- `module spider WGS` any size genome
- `module spider ALLPATHS-LG` any size genome
- `module spider Velvet` small genomes or lots of RAM
- `module spider SGA` any size genome

Resolving Assembly Gaps

Fragments and Gaps in Assemblies

- Repetitive regions can lead to a fragmented assembly
  - Transposable Elements
  - Ribosomal genes
  - Duplicated genes
  - Duplicated chromosomal regions
- Low (<40) or high (>60) %GC can result in reduced sequence coverage
  - Need adequate (>30x) sequence coverage for assembler to extend contigs with confidence
  - Can use specialized tools to resolve gaps resulting from low sequence coverage
    - `module spider GapFiller`
    - `module spider Opera`
    - `module spider SSPACE`

DNA Repetitive Regions

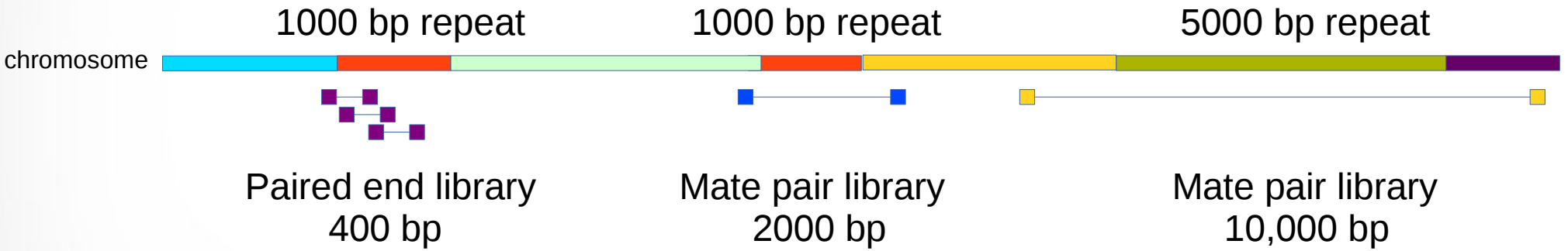


If a repeat is longer than both the read length and the distance between paired reads it becomes impossible to resolve the repeat region of the graph.

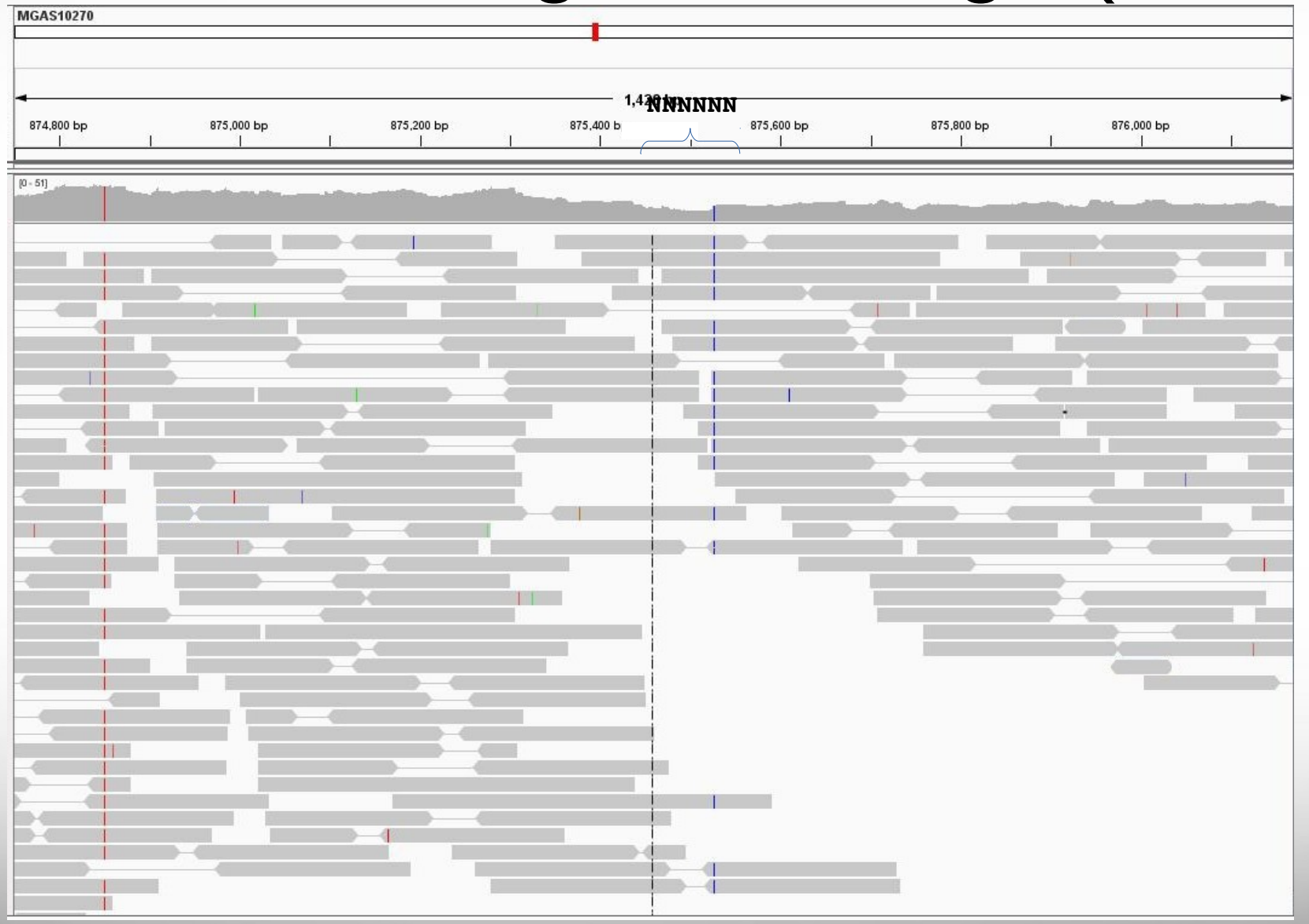
This means that even with a high coverage of error free reads, the result of an assembly will still be fragmented if long repeat regions are present.

White Paper: *de novo* Assembly in CLC Assembly Cell 4.0

Use Multiple Libraries to Resolve Repetitive Regions

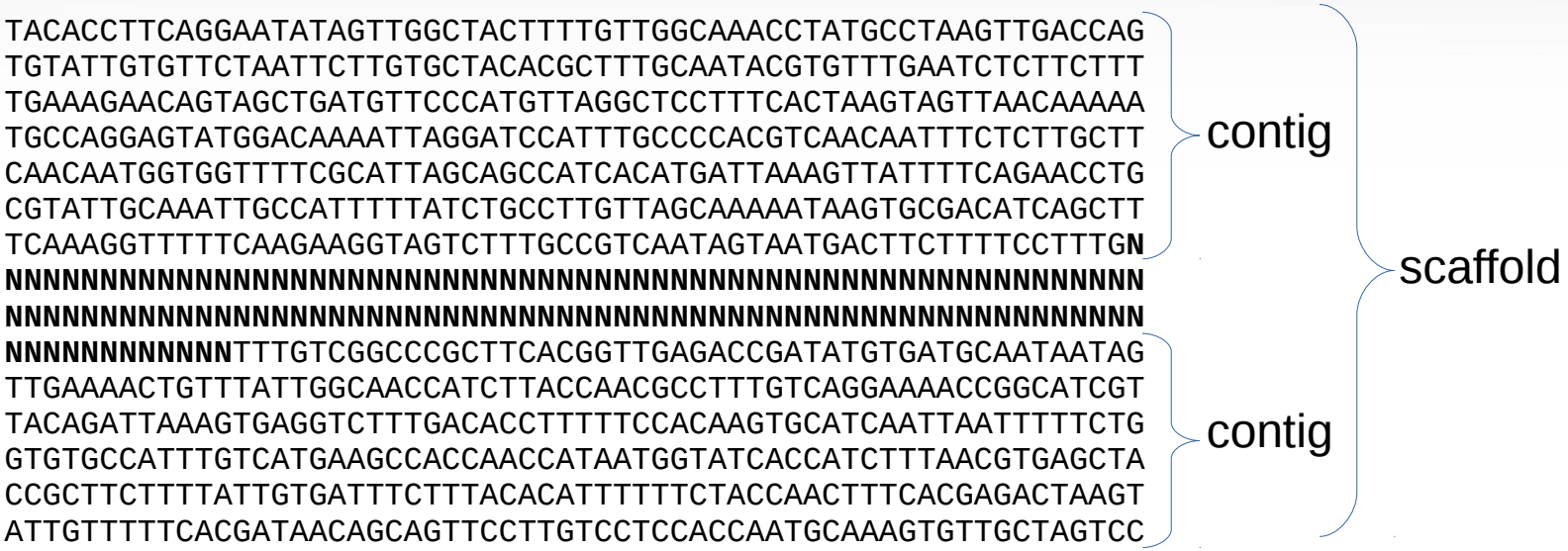


Example of Gap in a Region of Low Coverage Coverage (< 15x)



Gap Filling Results

134 Ns
before filling gap



121 nucleotides
after filling gap

Improve Draft Genome Assemblies

```
module spider Pilon
```

- Attempts improvements to the input genome
  - Single base differences
  - Small indels
  - Larger indel or block substitution events
  - Gap filling
  - Identification of local misassemblies, including optional opening of new gaps
- Input files
  - Genome fasta file
  - One or more bam alignment files

Other Assembly Considerations

- Reference Assisted Genome Assembly
- Merge multiple assemblies
  - `module spider Metassembler`
- Assemblathon 2 results

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33],

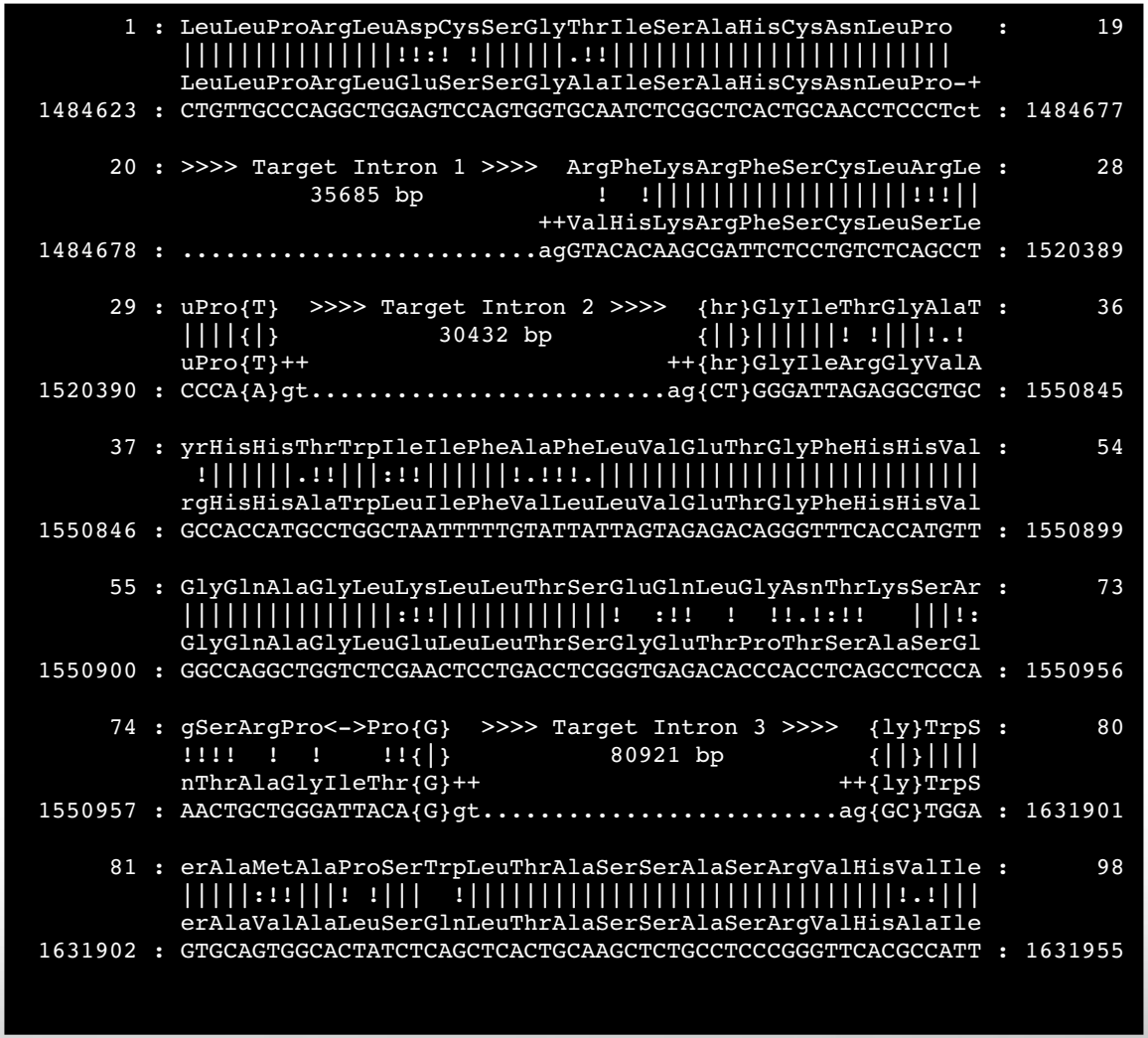
Genome Annotation

Annotating a *de novo* Assembly

- Align a set of proteins to *de novo* contigs/scaffolds
  - Use a set of proteins from a reference annotation
  - Use gene modeling tools to identify open reading frames
- Identify proteins that have amino acid changes
  - Non-synonymous (amino acid sequence change)
  - Frameshifts
  - Insertions/Deletions
  - Start/Stop gain/lost
  - Splice Sites

Computational Annotation of Contigs and Scaffolds

- Alignment tools
  - **module spider** Exonerate
    - protein to genome
      - shows splice sites, and introns (with lengths)
      - parse alignments coordinates with BioPerl
    - ESTs
  - RNA-seq
    - **module spider** GMAP-GSNAP
    - **module spider** STAR-STAR
    - **module spider** bowtie2
- Gene Prediction Tools
  - **module spider** geneid
  - **module spider** GeneMark-ES
  - **module spider** GeneMarkS
  - **module spider** SNAP-HMM
  - **module spider** Augustus
- Genome Annotation Pipeline
  - **module spider** MAKER



exonerate output protein to genome *H. sapiens*

Genome Assembly Completeness

The completeness of a genome can be estimated by using a set of highly conserved genes that are common to specific taxonomic groups

- BUSCO – uses single-copy genes to access genome assembly and annotation completeness
  - `module spider BUSCO`
  - evaluates % complete 'BUSCOs', % fragmented, % missing
- CEGMA – Core Eukaryotic Gene Mapping Approach: accesses a genome assembly for a small set of highly conserved genes.
  - Currently no further development planned
  - `module spider CEGMA`

Sequence Variant Calling

Sequence Variant Calling

- Start with aligning reads to a reference
  - GATK does not require QC trimming
  - Mark PCR duplicates with Picard
- Differentiate between sequencing errors and SNPs
  - Calling SNPs may require a min read depth of 10x (higher for indels)
  - Calling variants may require 1/3 of reads to contain SNP
  - Strand bias may result as a consequence of the sequencing chemistry's response to certain DNA sequence motifs but it can be detected computationally
- BLAST reads with SNPs to identify variant calls due to misalignments especially with duplicated genes
- Variant Call Format (vcf) – standard format of variant calls
- Identify multiple-nucleotide polymorphism (MNP)
  - Two SNPs within a single codon

	codon	
Reference:	**TTT**	Phe
SNP 1:	**TTA**	Leu
SNP 2:	**TAT**	Tyr
SNP 1 + 2:	**TAA**	**STOP**

Marking PCR Duplicates

- PCR duplicates are artifacts resulting from a PCR amplification step during NGS library preparations.
- PCR duplicates should be removed/marked as to not bias the frequency of variants or gene expression levels
  - Use picard tools to mark duplicates

```
module spider picard
```

Variant Calling Tools

Use bam file of sequence reads aligned to a reference as input for the following four work flows

- GATK

```
module spider GATK picard SAMtools
```

  - No need to QC trim reads, the GATK best practices pipeline will perform the necessary steps including marking PCR duplicates
  - You need a set of known variants for your species (dbSNP) or you can bootstrap your population to get variant frequency
  - Used in conjunction with other tools
    - samtools
    - picard
- SAMtools and BCFtools

```
module spider SAMtools BCFtools
```

- VarScan

```
module spider VarScan
```

- FreeBayes

```
module spider FreeBayes
```

Sample vcf File Format

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2         Sample3
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51  1|0:48:8:51,51  1/1:43:5:.,.
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50  0|1:3:5:65,3    0/0:41:3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:56,51  0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                   GT:GQ:DP    0/1:35:4        0/2:17:2        1/2:40:3
```

vcf File Column Descriptions

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51  1|0:48:8:51,51
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50  0|1:3:5:65,3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:56,51
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                   GT:GQ:DP    0/1:35:4        0/2:17:2
```

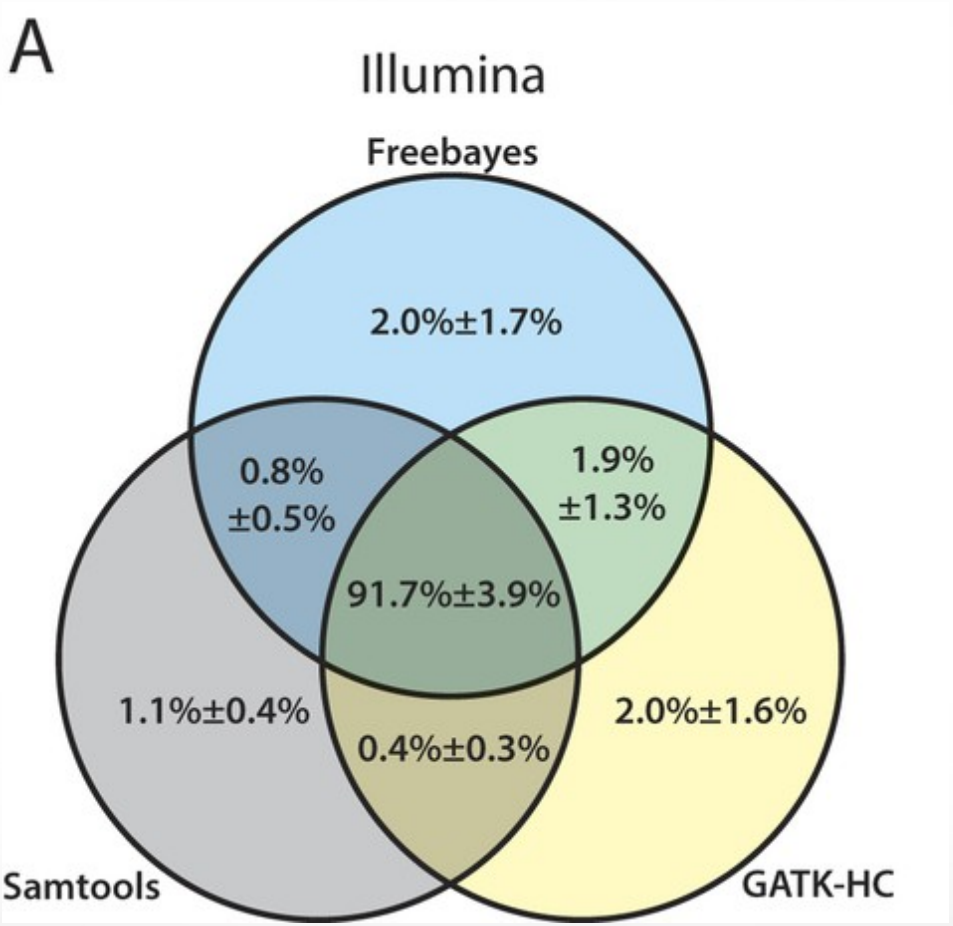
variants that are phased are inherited together

| indicates phased variants
/ indicates non-phased variants

Sample1 haplotypes: **GTGT** and **GTTT**
Sample2 haplotypes: **ATTT** and **GAGT**

https://www.broadinstitute.org/gatk/guide/tagged?tag=phasing

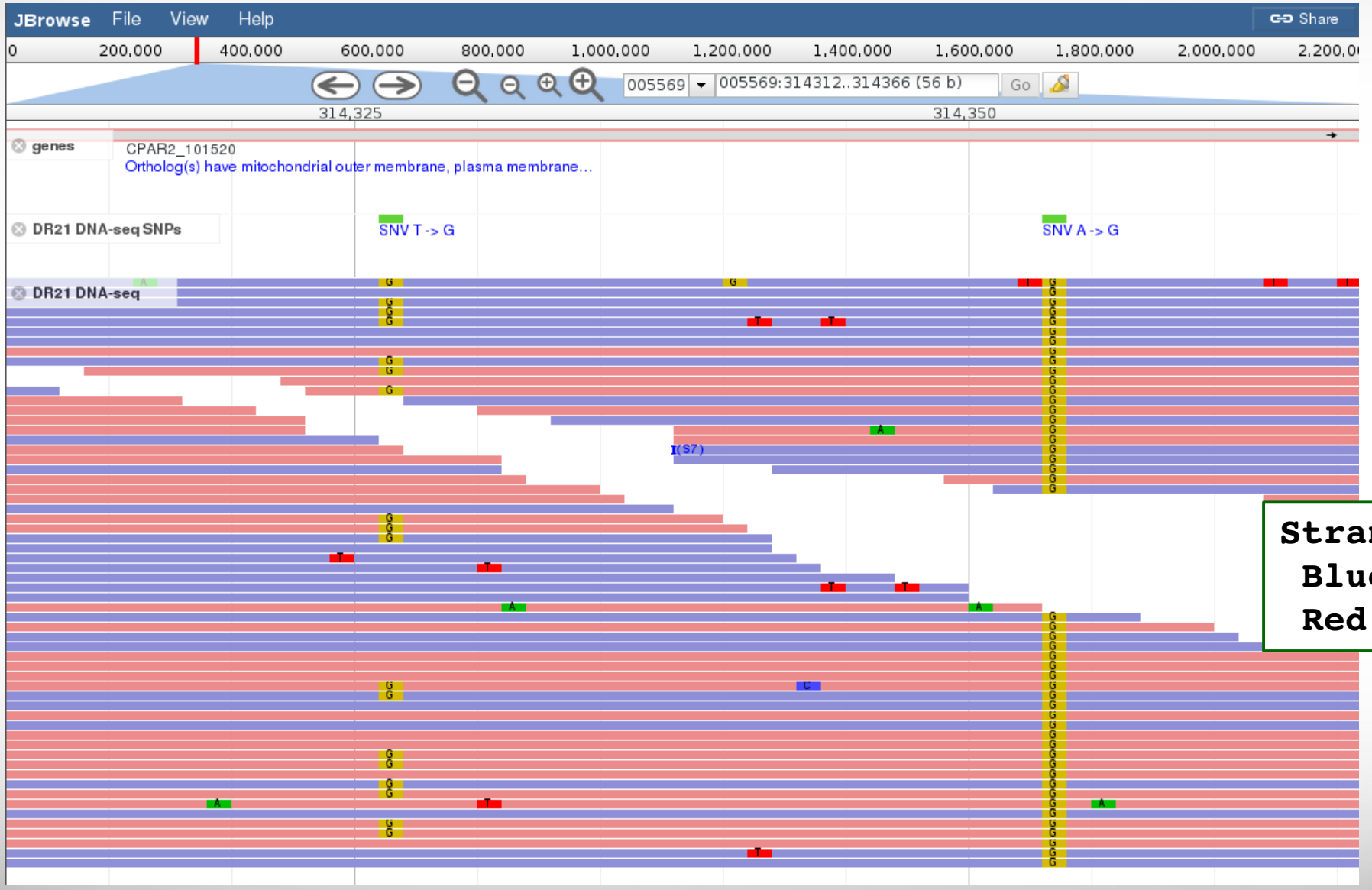
Summarizing Variant Calls from Different Tools



The mean percentage with standard deviation of confidence variant calls with equal to or higher than the quality score threshold of 20 are represented for (A) Illumina data sets

Huang et al 2015 doi:10.1038/srep17875

Viewing SNPs in a Diploid Organism



heterozygous SNP

homozygous SNP

Consequence of Amino Acid Change

- Assess consequence of amino acid change based on sequence conservation across multiple species using the PROVEAN tool
- Variants with a score equal to or below -2.5 are considered “deleterious”

```
module spider PROVEAN
```

```
## PROVEAN v1.1 output ##
# Query sequence file:  CTRG_00013.fa
# Variation file:   CTRG_00013.var
# Protein database: /scratch/datasets/blast/nr
[16:01:13] searching related sequences...
[16:16:36] clustering subject sequences...
# Number of clusters:   30
# Number of supporting sequences used: 245
[16:18:39] computing delta alignment scores...
## PROVEAN scores ##
# VARIATION SCORE
A431S   -0.455
E411K   -3.051
E226Q   -1.564
```

Verify enough supporting sequences found

“deleterious”

Annotate Variants

- A file of variant calls in vcf format is needed
- A reference sequence with gene annotations is needed
- snpEff annotates a vcf file `module spider snpEff`
  - There are > 2,500 pre-built databases available
  - Annotates MNP (multiple nucleotide polymorphism)
    - Codon change due to two SNPs: ACA → GGA

```
5       325795  .       AC      GG      23.8901 .
AB=0.428571;ABP=3.32051;AC=1;AF=0.5;AN=2;AO=3;CIGAR=2X;DP=7;DPB=7;DPRA=0;EPP=3.73412;
EPPR=3.0103;GTI=0;LEN=2;MEANALT=1;MQM=33;MQMR=48.5;NS=1;NUMALT=1;ODDS=5.49681;PAIRED=0;
PAIREDR=0.5;PAO=0;PQA=0;PQR=0;PRO=0;QA=114;QR=150;RO=4;RPL=3;RPP=9.52472;RPPR=3.0103;
RPR=0;RUN=1;SAF=2;SAP=3.73412;SAR=1;SRF=2;SRP=3.0103;SRR=2;TYPE=mnp;technology.ILLUMINA=1;
ANN=GG|missense_variant|MODERATE|CD36_51230|CD36_51230|transcript|CAX41505.1|
protein_coding|1/1|c.1657_1658delACinsGG|p.Thr553Gly|1657/1851|1657/1851|553/616||
GT:DP:RO:QR:AO:QA:GL    0/1:7:4:150:3:114:-6.7054,0,-11.1847
```

Example of Sequencing Strand Bias

Strand:
Blue = +
Red = -

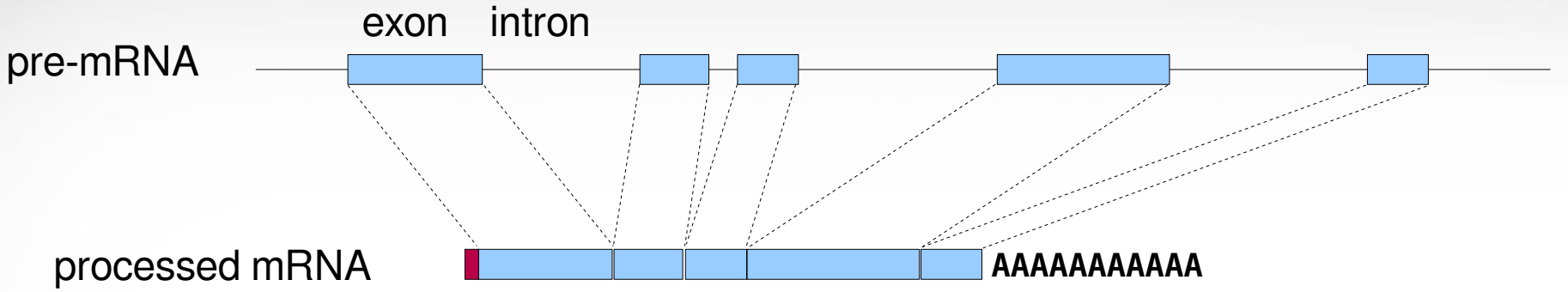
Identify/estimate strand bias using values in vcf file

Strand bias counts:
SRF, SRR, SAF, SAR
Bias estimate:
SAP

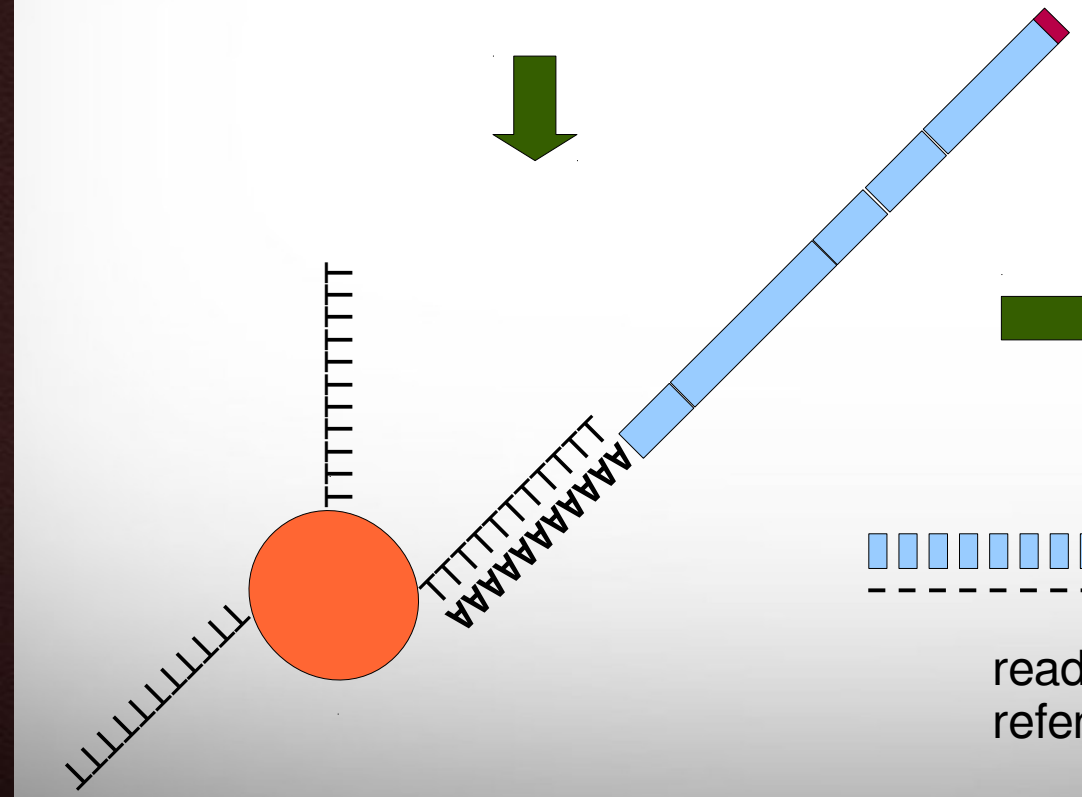
IonTorrent Proton

RNA-seq

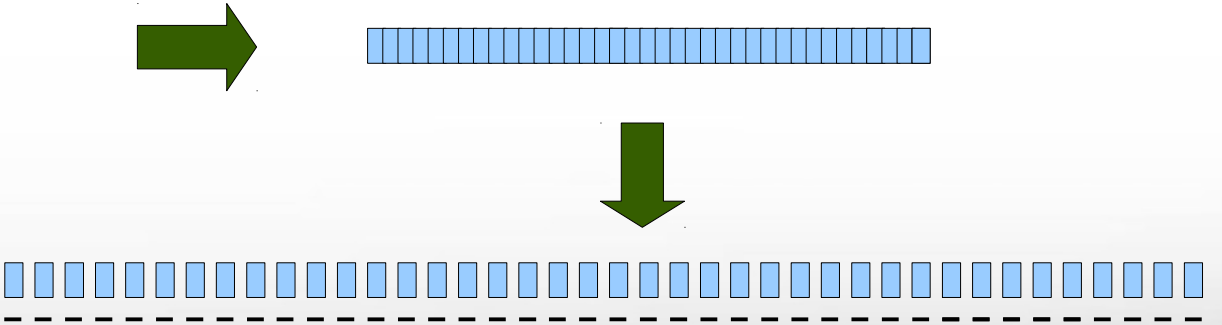
Example of RNA sequencing



mRNA strands are captured by their Poly(A) tail using Poly(T) coated magnetic beads



exon region is fragmented and sequenced



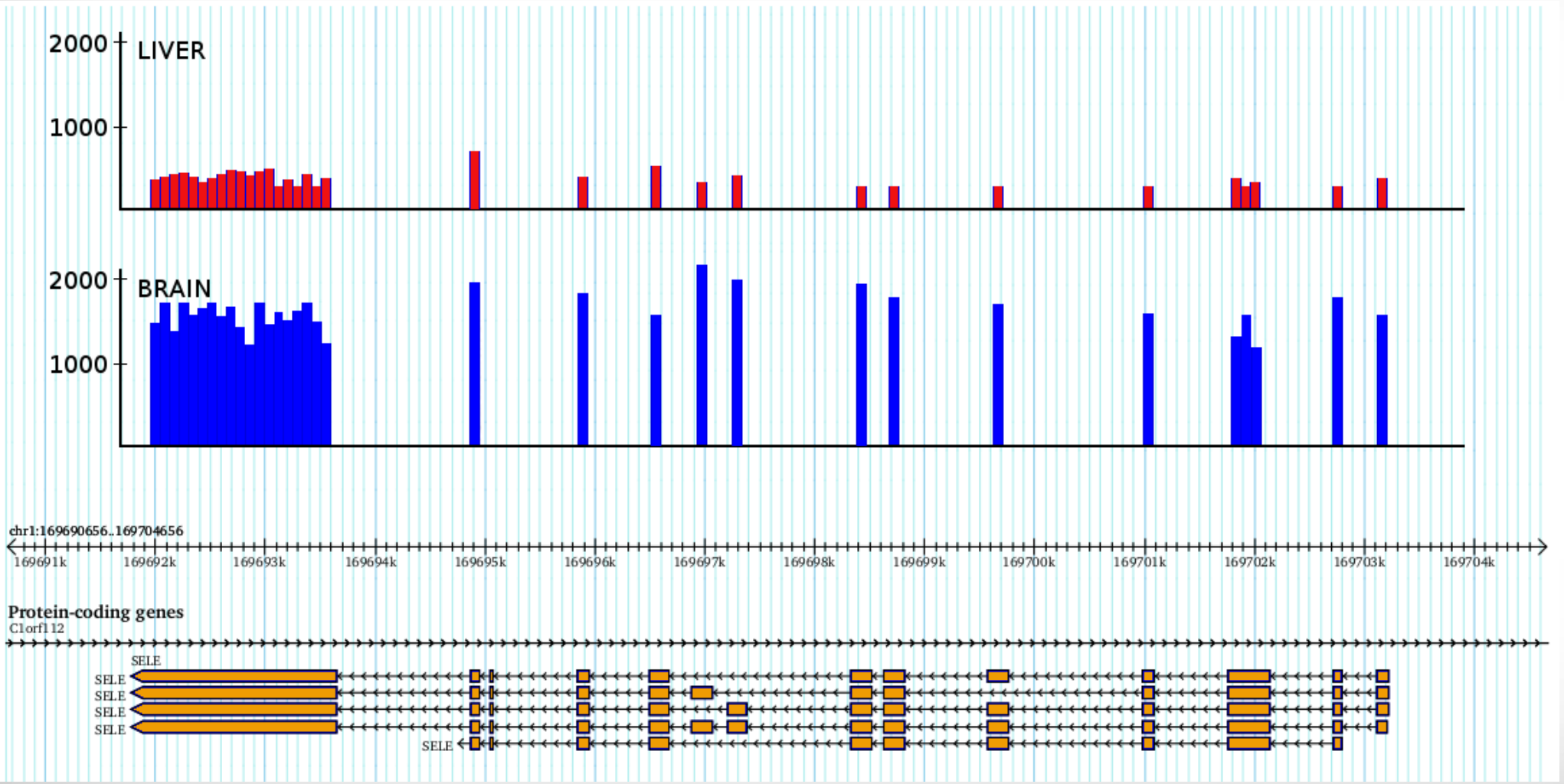
reads are then mapped to a reference genome or reference sequences or assembled

RNA-seq Applications

- Transcriptome assembly (find isoforms and rare transcripts)
  - *de novo* (Trinity, SOAPdenovo-Trans)
  - reference based (Trinity, StringTie)
- Differential Expression (DE)
  - Bowtie, TopHat, Cufflinks, Cuffmerge, Cuffdiff
  - DESeq and DESeq2 (R)
  - EdgeR (R)
- Variant Calling
  - STAR/Picard/GATK (Haplotype Caller (HC) in RNA-seq mode)
- Genome Annotation
  - Align to assembly for validation of gene models
- *de novo* genome assembly scaffolding
  - L_RNA_scaffolder

RNA-seq for Differential Expression

RNA-seq Differential Expression (DE)



http://www.illumina.com/technology/mrna_seq.ilmn

RNA-seq Transcript Alignment Counting

- Alignment based
  - Non-normalized alignment counts
    - HTSeq-count
  - Normalized (RPKM, FPKM, TPM)
    - eXpress (outputs FPKM)
    - RSEM (isoform/gene level estimates without RPKM or FPKM)
    - Trinity Transcript Quantification
      - A Trinity script can run: Kallisto, RSEM, eXpress, Salmon
- Non-Alignment based
  - Kallisto (pseudoalignment)
  - Salmon (lightweight alignment)
  - Sailfish (k-mer)

RPKM vs FPKM vs TPM

- The number of **R**eads **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for single end reads
- The number of **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for paired-end reads
    - If both paired reads align to a transcript then they are counted as one alignment
- **T**ranscripts **P**er kilobase **M**illion
  - Normalize for gene length first
  - Normalize for sequence depth second

http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Sailfish

- Transcript quantification from RNA-seq data
- Requires a set of target transcripts (fasta)
  - From a reference or a *de-novo* assembly
- Requires sequence reads (fasta or fastq)
- Sample output quant.sf file:

Name	Length	EffectiveLength	TPM	NumReads
TRINITY_DN30_c0_g1_i1	215	68.4635	236.773	233
TRINITY_DN43_c0_g1_i1	280	102.34	5971.5	8784
TRINITY_DN88_c0_g1_i1	217	69.3036	191.74	191
TRINITY_DN59_c0_g1_i1	393	194.337	4092.64	11432
TRINITY_DN98_c0_g1_i1	205	64.4299	1097.09	1016
TRINITY_DN17_c0_g1_i1	310	122.99	2634.35	4657

Differential Expression (DE) based on alignment counts

Non-normalized abundance counts are used as input for DE analysis

- DESeq2
  - DE for genes not isoforms
- edgeR
  - DE at gene, exon (isoform) or transcript level
- EBSeq
  - DE for isoforms
- DEXSeq
  - DEU differential exon usage

Tuxedo Suite

- TopHat – splice aware alignment of RNA-seq reads using Bowtie2
  - TopHat is superseded by HISAT2
- Cufflinks – assembles aligned reads into transcripts and estimates their abundances
- Cuffdiff – compares RNA-seq abundance (expression) levels of two samples or groups

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
CAWT_00001	CAWG_00001	-	chr_1.1:8373-9093	q1	q2	OK	111.944	163.869	0.549763	0.768107	0.58795	0.996768	no
CAWT_00002	CAWG_00002	-	chr_1.1:11447-12425	q1	q2	OK	14.5992	30.9037	1.08189	1.3841	0.2921	0.98312	no
CAWT_00003	CAWG_00003	-	chr_1.1:14130-14451	q1	q2	OK	248.323	259.152	0.0615814	0.172186	0.94685	0.996768	no
CAWT_00004	CAWG_00004	-	chr_1.1:14890-16045	q1	q2	OK	60.9546	86.0009	0.496617	0.604904	0.6204	0.996768	no
…													
…													
CAWT_01628	CAWG_01628	-	chr1.2:664522-665344	q1	q2	OK	3.56447	157.849	5.46871	6.64693	0.00015	0.0482417	yes

p_value = The uncorrected p-value of the test statistic.
q_value = The FDR-adjusted p-value of the test statistic

R Bioconductor

- Bioconductor packages can be found in this R version

```
module load R_tamu/3.3.1-intel-2015B-default-mt
```

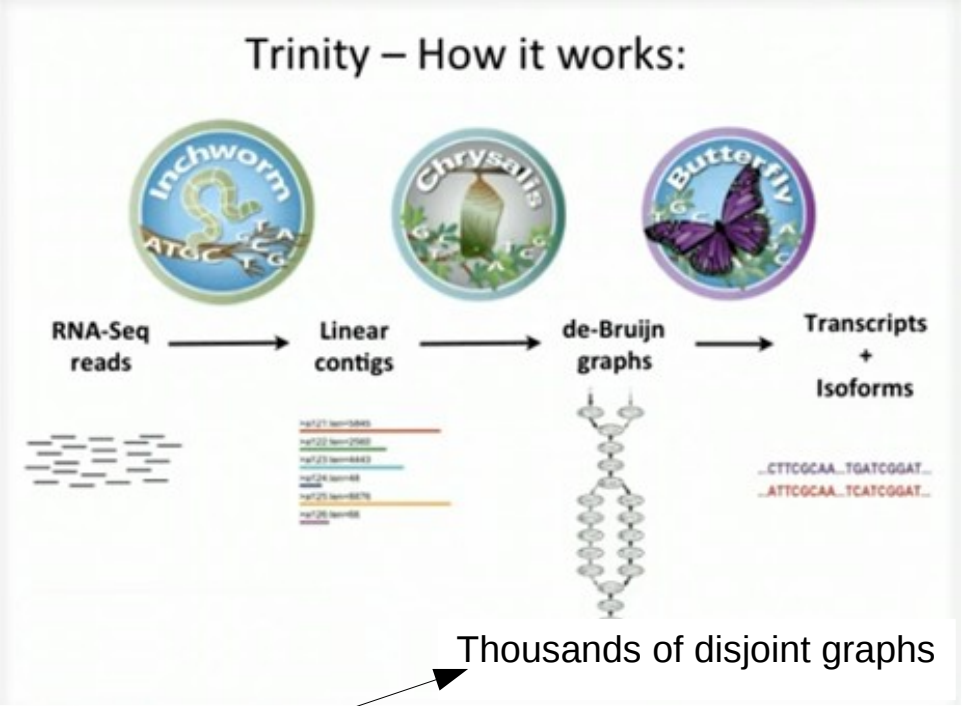
- Popular R bioconductor packages for RNA-seq
  - CQN – Normalization of RNA-seq data
  - edgeR – Differential gene expression
  - DESeq, DESeq2 – Differential gene expression
  - cummeRbund – analysis/visualization of cufflinks data

RNA-seq for Transcriptome Assembly

RNA-seq Transcriptome Assembly

- Assembly with a reference genome
  - `module spider Trinity`
  - `module spider Bowtie2 Tophat Cufflinks`
  - `module spider Scripture`
- *de novo a*ssembly without a reference genome
  - `module spider Trinity`

Trinity – How it works:



Thousands of disjoint graphs

Broad Institute

ideally one graph per gene

http://www.rna-seqblog.com/a-collection-of-new-rna-seq-videos-from-the-broad-institute/

Running Trinity on Ada

- Trinity uses 100,000s of intermediate files
  - Contact **help@hprc.tamu.edu** and request a file quota increase before running Trinity
  - Run one Trinity job at a time and check resource usage
    - `showquota`
    - It is recommended not to run multiple Trinity jobs unless you know memory usage and an estimate of the number of temporary files generated
  - Trinity creates checkpoints and can be restarted if it stops (but not in Galaxy)
- See GCATemplates for sample Trinity commands

Running Trinity on Ada 64GB node

- Use all cores and memory on a node
  - 54GB avail memory on 64GB nodes

```
#BSUB -n 20
#BSUB -R "span[ptile=20]"
#BSUB -R "rusage[mem=2700]"
#BSUB -M 2700
```

- Recommended Trinity options

```
--max_memory 53G
--CPU 20
--inchworm_cpu 6
--no_version_check
```

Running Trinity on Ada 256GB node

- Use all cores and memory on a node
  - ~246GB avail memory on 256GB nodes

```
#BSUB -n 20
#BSUB -R "span[ptile=20]"
#BSUB -R "rusage[mem=12300]"
#BSUB -M 12300
#BSUB -R "select[mem256gb]"
```

- Recommended Trinity options

```
--max_memory 245G
--CPU 20
--inchworm_cpu 6
--no_version_check
```

Running Trinity on Ada 1TB node

- Use all cores and memory on a node
  - 1TB avail memory on 1TB nodes

```
#BSUB -n 40
#BSUB -R "span[ptile=40]"
#BSUB -R "rusage[mem=25000]"
#BSUB -M 25000
#BSUB -q xlarge
#BSUB -R "select[mem1tb]"
```

- Recommended Trinity options

```
--max_memory 995G
--CPU 40
--inchworm_cpu 6
--no_version_check
```

RNA-seq Transcriptome Assembly Evaluation

```
module spider DETONATE
```

- Input transcriptome fasta assembly and sequence reads fastq or fasta files
- RSEM-EVAL used for reference-free evaluation
- REF-EVAL used for reference-based evaluation
- Higher score = better evaluation
- Sample output:

```
Score                        -30198099.46
Number_of_contigs              1976
Number_of_alignable_reads                  1140584
Number_of_alignments_in_total              1434453
```

RNA-seq Transcriptome Assembly Annotation

```
module spider Trinotate
```

- Will run a series of tools to annotate an assembly
  - RNAMMER
    - predicts 5s/8s, 16s/18s, 23s/28s ribosomal RNA
  - TransDecoder
    - predicts coding regions
  - BLAST+ (SwissProt db)
  - HMMER (PFAM db)
  - SignalP
    - predicts presence and location of signal peptide cleavage sites in amino acid sequences
  - tmhmm
    - prediction of transmembrane helices in proteins
- Results are saved in SQLite db and as a summary file: Trinotate.xls

rna-seqblog.com

DON'T MISS Maverix Biomics Offers Free Trial of Self-Service RNA-Seq Data Analysis

Search...

RNA-Seq Blog
Transcriptome Sequencing Research & Industry News

HOME | NEWS » | EVENTS » | JOBS » | TECHNOLOGY » | DATA ANALYSIS » | BLOG | CONTACT »

Tag Archives: TMM

Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data

October 30, 2015 0 1,501 Views

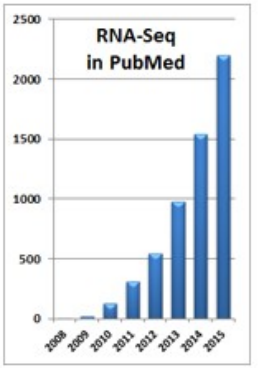
Recently, rapid improvements in technology and decrease in sequencing costs have made RNA-Seq a widely used technique to quantify gene expression levels. Various normalization approaches have been proposed, owing to the importance of normalization in the analysis of RNA-Seq data. …

Read More »

An iteration normalization and test method for differential expression analysis of RNA-seq data

STAY CONNECTED

PUBLICATIONS TREND



RECENT RNA-SEQ PUBS

RED: A Java-MySQL Software for Identifying

SUBSCRIBE TO THE RNA-SEQ BLOG

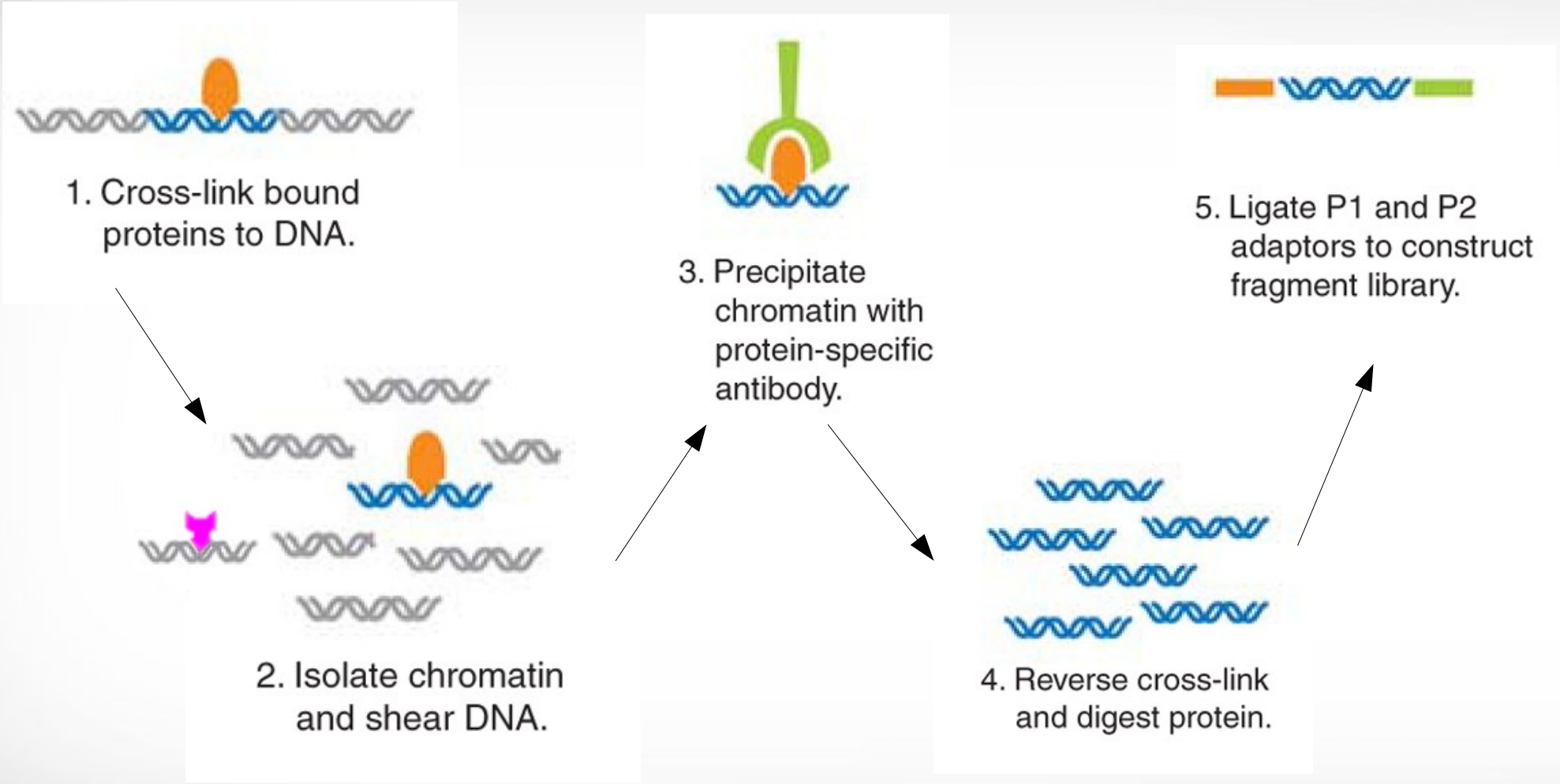
email address

Subscribe

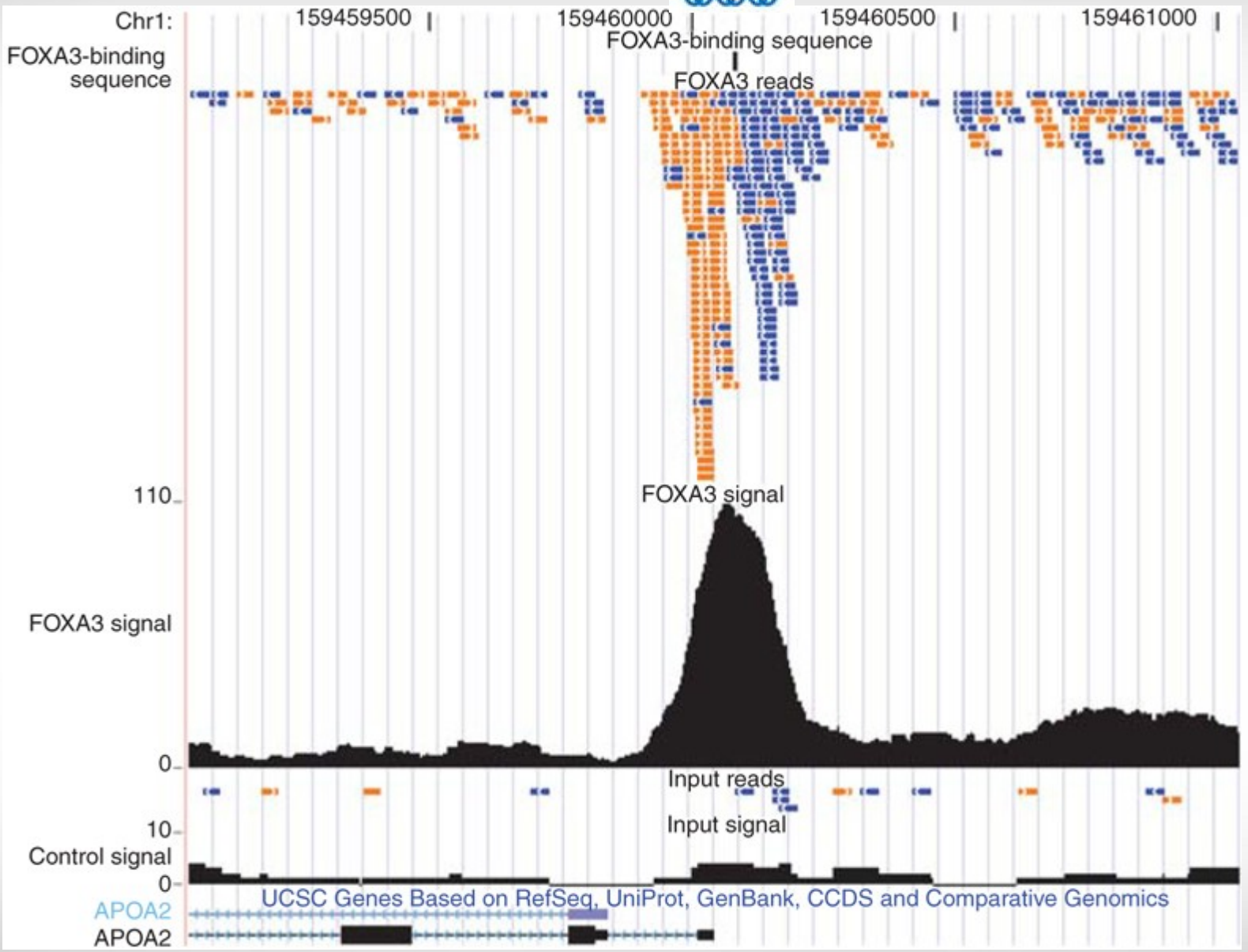
RNA-SEQ PRODUCTS & SERVICES

ChIP-seq

Chromatin immunoprecipitation (ChIP) is a technique for identifying and characterizing elements in protein-DNA interactions involved in gene regulation or chromatin organization.

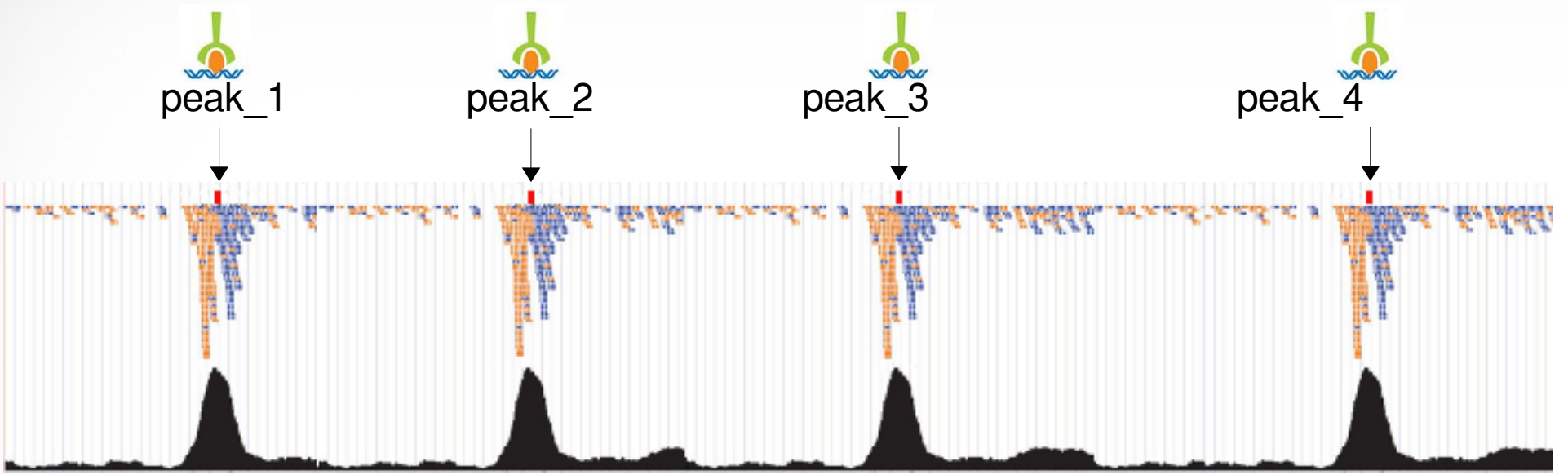


Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system
Nature Methods 6, (2009)



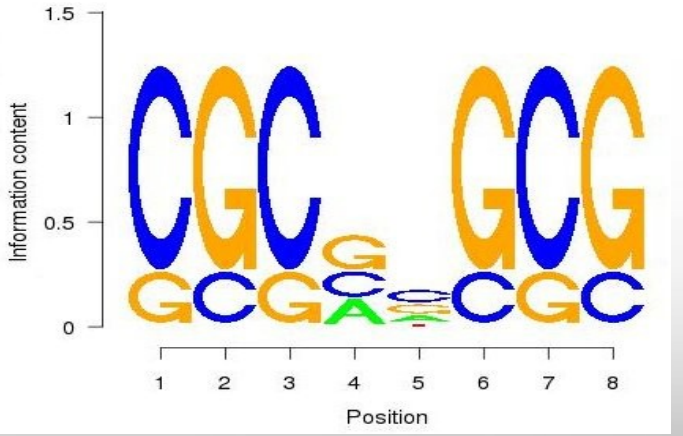
Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system
Nature Methods 6, (2009)

The goal is to find a consensus DNA sequence among the sequences at each peak which will give us the DNA sequence motif that a protein recognizes and binds



A sequence logo can be used to represent the DNA sequence motif where the protein binds

Generate a sequence logo with the R package seqLogo



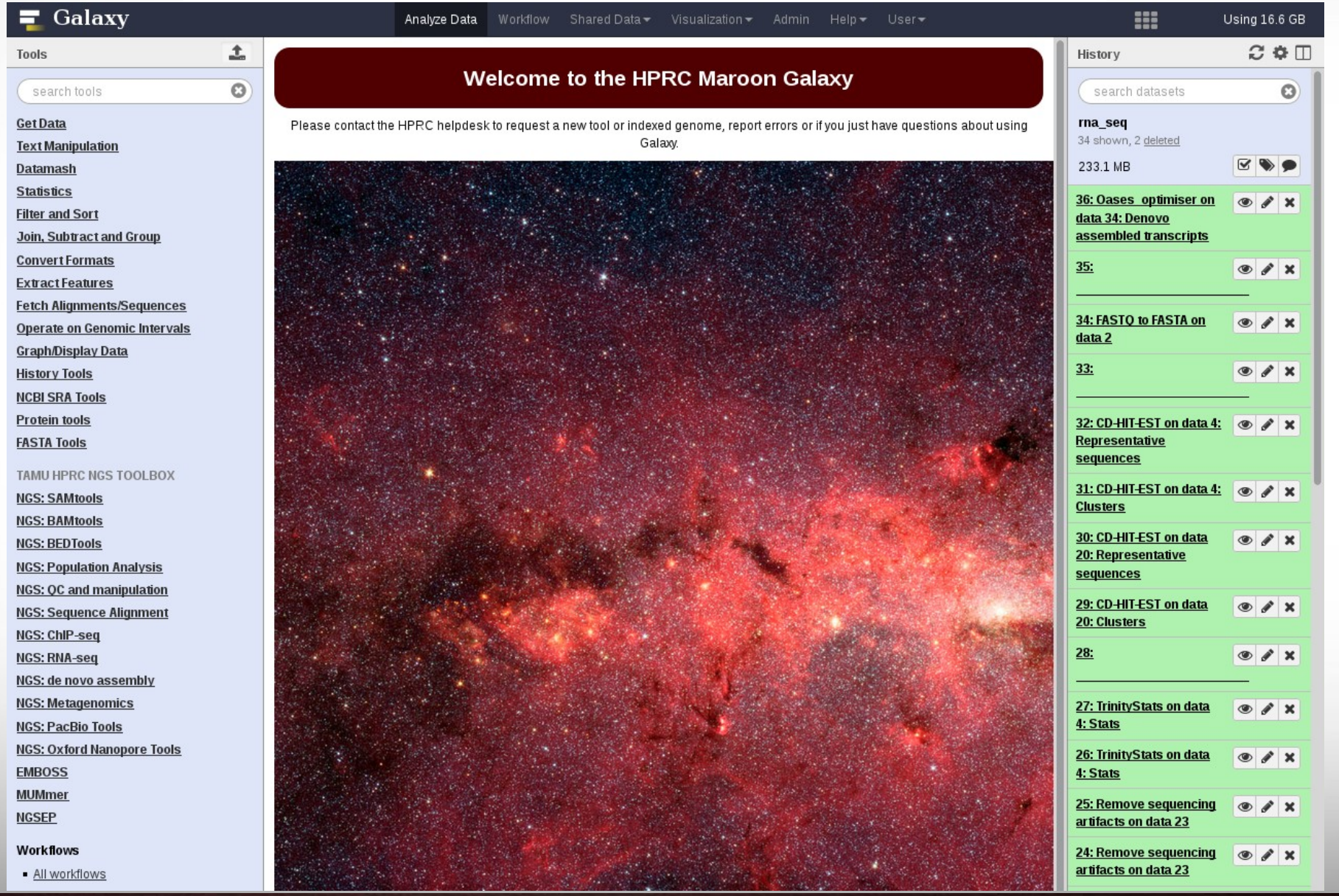

```
module load R_tamu/3.3.1-intel-2015B-default-mt
```

ChIP-seq Tools

- Protein-DNA interactions
  - `module spider MACS`
  - `module spider MACS2`
- Subdivision of ChIP-seq regions into discrete signal peaks
  - `module spider PeakSplitter`
- Peak caller
  - `module spider PeakRanger`
  - `module spider BroadPeak`
- Identify enriched domains from histone modification ChIP-seq data
  - `module spider SICER`

Galaxy on Ada

https://hprcgalaxy.tamu.edu/maroon/



Galaxy Notes on Ada

https://hprc.tamu.edu/wiki/index.php/Ada:Galaxy

Galaxy

Contents [hide]

Account Security

Do not share your Galaxy account with anyone. Galaxy uses the TAMU Central Authentication Service which is linked to your TAMU account.

Make sure you always logout of Galaxy by selecting User -> Logout and then click the Logout button on the next screen and then close your browser when you are finished using Galaxy.

FishCamp Galaxy Accounts

The FishCamp Galaxy instance is reserved for training purposes such as Galaxy workshops.

When requesting access to FishCamp for a training workshop, please include your ada NetID in your request.

- The FishCamp Galaxy is configured for training purposes.
  - Most jobs will run a maximum of 1 hour.
  - This is to enable jobs to be scheduled faster in the cluster queue.
  - Keep your input datasets small so that they will complete within one hour.

FishCamp Galaxy is not intended for research projects and data on FishCamp Galaxy should be considered to have short term accessibility.

Request a Maroon Galaxy account only if you have data to analyze.

The tools available on FishCamp and Maroon are the same.

HPRC Resources

- Free Help
  - Send an email to help@hprc.tamu.edu if you have any questions regarding Bioinformatics tools usage on HPRC clusters
    - First spend some time investigating the error
      - read log files, stdout file, stderr file, tool manual
      - Google search
      - Google user groups
    - Include details about your issue
      - Which cluster or which Galaxy you are using
      - Which tool you are using
      - Which modules you have loaded
      - Commands you used in your job script
      - Error messages you are seeing
- HPRC NGS data analysis tools Documentation
  - https://hprc.tamu.edu/wiki/index.php/Ada:Bioinformatics

Upcoming HPRC Short Courses

Topics	Date/Time
Intermediate Scripting	3-5pm Fri, Mar 24
Introduction to Code Parallelization using OpenMP	3-5pm, Wed, Mar 29
Jetstream: Hands-on Workshop	1:30- 4:00pm Fri, Mar 31
Introduction to Code Parallelization using MPI	3-5 pm, Wed, April 5
Data Security and Ethics	3-5pm Wed, April 7
Introduction to Atomistic Simulations	3-5pm Wed, April 12
Introduction to FORTRAN	3-5pm Wed, April 19
Visualization Portal	3-5pm Fri, April 21
Databases	3-5pm Wed, April 26

Register or see the full list of short courses

https://hprc.tamu.edu/register/classlist.php

Thank you.

Any question?